CORE
Metadata, citation and similar papers at core.ac.uk
Provided by OpenCommons at University of Connecticut

University of Connecticut
OpenCommons@UConn

Doctoral Dissertations
University of Connecticut Graduate School

4-26-2017

# The Impact of Online Training on the Reliability of Direct Behavior Ratings

Nicholas J. Crovello
*University of Connecticut*, nicholas.crovello@uconn.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation
Crovello, Nicholas J., "The Impact of Online Training on the Reliability of Direct Behavior Ratings" (2017). *Doctoral Dissertations*. 1403.
https://opencommons.uconn.edu/dissertations/1403

# The Impact of Online Training on the Reliability of Direct Behavior Ratings

Nicholas J. Crovello, PhD

University of Connecticut, 2017

Direct Behavior Rating (DBR) has emerged as a useful assessment method to identify behavioral risk and monitor behavioral progress. The development of an online training module has provided an option for low cost training accessible to a wide variety of users at their convenience. Although evaluation of the DBR online training module has demonstrated improved rater accuracy following completion, reliability of obtained scores has not been fully explored. An improvement in reliability of scores is desired to allow for flexible use across raters and increased efficiency in decision-making. In this study, four teachers simultaneously rated the activity engagement and disruptive behavior of six children during their ice skating activity at a summer day camp. Ratings were analyzed within a generalizability theory (GT) framework. Results suggest that completion of the online training module may result in some improvement in the reliability of data generated from DBR-SIS.

# The Impact of Online Training on the Reliability of Direct Behavior Ratings

Nicholas J. Crovello

B.A., Siena College, 2013

M.A., University of Connecticut, 2014

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of Doctor of Philosophy

at the

University of Connecticut

2017

Copyright by

Nicholas J. Crovello

2017

APPROVAL PAGE

Doctor of Philosophy Dissertation

The Impact of Online Training on the Reliability of Direct Behavior Ratings

Presented by

Nicholas J. Crovello, B.A., M.A.

Major Advisor _____

Sandra M. Chafouleas

Associate Advisor _____

Hariharan Swaminathan

Associate Advisor_____

Austin H. Johnson

University of Connecticut

2017

Acknowledgements

I would like to begin by thanking my major advisor, Sandra Chafouleas, without whom this success would not be possible. Through her dedication, motivation, and expertise, she has fostered personal and professional growth since the day I arrived on campus. However, I would be remiss not to recognize the significant contributions of Hariharan Swaminathan, Austin Johnson, and the entirety of the University of Connecticut faculty in helping to shape my identity as a researcher and a practitioner.

I would also like to thank my family for impressing on me the importance of education and allowing me the opportunity to pursue it at the highest level. I am very grateful for my participants, as their enthusiasm and devotion was critical to successful completion of this study. Finally, this work is dedicated to Kerry, whose unconditional love and support is unparalleled.

Table of Contents

List of Tables and Figures

In the current educational climate, schools are increasingly responsible for the behavioral outcomes of their students. According to the Every Student Succeeds Act (ESSA; 2015), schools are urged to consider "implementation of a school-wide tiered model to prevent and address problem behavior…coordinated with similar activities and services carried out under IDEA." Increased emphasis on behavior in the legislature is likely due to the substantial prevalence of behavior disorders (Merikangas et al., 2010) and the growing amount of resources schools are required to devote to address behavioral issues (Scholastic, 2012). Thus, in order to comply with federal regulations and to deliver positive behavior outcomes for both individual students and the complete school system, research has advanced the use of a structured decision making model (i.e. multi-tiered systems of support; MTSS). MTSS is essentially a framework that outlines the types of supports a student requires given a specified level of need. MTSS relies on the implementation of high quality, evidence based interventions delivered with fidelity across different degrees of support intensity: universal, targeted, and individualized (Gresham, 2011). Universal interventions are provided to all students and are expected to be effective for those who need a low level of additional support (Tilly, 2008). It is the level at which school wide teaching practices and policies are implemented to prevent occurrences of problem behavior. At the targeted level, additional supports are implemented for students who are at-risk for or demonstrating non-severe levels of problem behavior. Finally, students who exhibit the most significant behavioral challenges are supported through intensive, individualized interventions.

According to the National Association of School Psychologists, regulations in ESSA compel educators to develop assessments as a means to "inform data-based decisions made about the needs of individual students and the school system as a whole" (NASP, 2016). Development of these assessment methods is crucial as valid, reliable, and accurate data are needed at each tier

in order to (a) appropriately determine the intensity of support that a child should receive and (b) evaluate the effectiveness of supports currently being provided (Christ, Riley-Tillman, & Chafouleas, 2009; Gresham, 2011). More specifically, behavior assessment within an MTSS model must strive to correctly determine which students demonstrate behavioral risk (screening) and to evaluate a student's response to an intervention (progress monitoring).

In order to best inform progress monitoring and screening decisions within an MTSS model, data derived from a behavior assessment tool must be defensible, flexible, efficient, and repeatable (Chafouleas, Riley-Tillman, & Christ, 2009; Christ et al., 2009). That is, data from these tools must have sound psychometric qualities, be able to be used in a variety of settings and for multiple purposes, and be able to be completed repeatedly within a reasonable amount of time. Across the growing literature of behavior assessment, Direct Behavior Rating (DBR) has emerged as meeting the requirements (defensible, flexible, efficient, and repeatable) for assessing behavior within an MTSS model (Chafouleas et al., 2009; Christ et al., 2009).

**Direct Behavior Rating**

The roots of DBR lie in informal home-school communication methods such as home-school notes and daily progress reports (Kelley, 1990; Chafouleas, Riley-Tillman, & MacDougal, 2002). DBR can be best described as a hybrid tool that utilizes aspects of both direct observation and a rating scale. In using DBR, a rater makes an estimate of the percentage of time a student was engaged in a given behavior following a pre-specified observation period (Chafouleas, 2011). Although multi-item scales have been explored, the majority of the research regarding the psychometric properties of DBR has focused on Single Item Scales (DBR-SIS; Volpe & Briesch, 2012). DBR-SIS most often involves a rating of the three core behavioral competencies that are conceptualized as relevant to classroom success: academically engaged,

respectful, and disruptive (see Appendix A for an example DBR form with operational definitions for each core behavior). Each behavior is defined globally, meaning that the definition is comprehensive for all specific actions that comprise the overall behavior. Global definitions (i.e. academically engaged) were chosen over specific definitions (e.g., hand raising) given evidence of improved rating accuracy (Riley-Tillman, Chafouleas, Christ, Briesch, & LeBel, 2009). Following the development and instrumentation of the DBR scale, research turned to evaluating its capacity to be an effective tool within an MTSS model.

**Progress monitoring.** As previously mentioned, progress monitoring (formative assessment) is one of the cornerstones of an RTI problem solving model. Initial research on DBR-SIS focused on validating its use in formative assessment (Christ et al., 2009; Riley-Tillman, Methe, & Weegar, 2009). In determining whether data from a particular tool is useful in formative assessment, it is necessary to evaluate that tool's sensitivity to behavioral change. Riley-Tillman and colleagues (2008) detected similar trends and consistency when comparing data collected using DBR to data collected using a systematic direct observation (SDO) technique. Similarly, more recent investigations have supported the sensitivity of DBR to behavioral change across diverse settings (Chafouleas, Sanetti, Kilgus, & Maggin, 2012).

**Screening**. In order to create a more efficient assessment method, it is desirable that that tool serves a dual capacity in both progress monitoring and screening (Glover & Albers, 2007; Chafouleas, Kilgus, Jaffery, Riley-Tillman, Welsh, & Christ, 2013). In addition, this dual capacity allows a tool to have flexible use across reasons for assessment. Thus, recent research has turned towards the validation of DBR-SIS for use in screening (Christ et al., 2009). The overarching goal of screening is to correctly identify students at risk for demonstrating behavioral problems. Therefore, a valid screening tool is sensitive enough to identify those

students who are at risk, but specific enough to discriminate between those who are truly at risk and those who are not. In other words, the ideal screening tool maximizes the rates of results that are either true positive or true negative. A major line of research has emerged regarding the capacity for DBR to act as a sensitive, specific, and efficient tool for screening behavior (Kilgus, Chafouleas, Riley-Tillman, & Welsh, 2012; Chafouleas et al., 2013; Kilgus, Riley-Tillman, Chafouleas, Christ, & Welsh, 2014). Within these studies, the authors established optimal cut scores for determining at-risk status. It is important to note that the cut scores varied by grade level grouping and that disruptive tended to carry more importance in lower grades whereas academically engaged tended to carry more importance in higher grades (Chafouleas et al., 2013; Kilgus et al., 2014). Given the evidence to support DBR-SIS as a valid tool for use in both progress monitoring and screening, it is logical to explore methods to improve the validity, accuracy, and reliability of data generated from DBR-SIS.

**Rater Training**

Although DBR-SIS has demonstrated technical adequacy and utility as both a screener and a progress monitor, it is not without its limitations given evidence of rater error (Chafouleas et al., 2015), which has been traditionally operationalized as a lack of consistency between two scores (Cone, 1977). Error (i.e. the deviation of an observed value from its expected value) is an inherent part of any measurement technique, regardless of the method that is used. Error, however, may be more pronounced when raters assign a rating based on their subjective perception of a behavior, such as they do when using DBR. Raters can be inaccurate or evidence poor reliability (i.e. differently rate the same behavior under the same conditions; Chafouleas, Riley-Tillman, Jaffery, Miller, & Harrison, 2015; Chafouleas, Briesch, Riley-Tillman, Christ, Black, & Kilgus, 2010). Given any assessment instrument, inaccurate ratings or unreliable data

can be due to a variety of factors. Cronbach and colleagues (1972) suggested that factors that may cause ratings to be inaccurate include rater characteristics, complexity of the observation system, and conditions of the observation. However, researchers have explored a variety of techniques to mitigate the factors that may lead to error.

Rater training has been suggested as a means to improve rater accuracy and reliability (Stamoulis & Hauenstein, 1993). To this point, a series of investigations have been conducted in order to determine the most efficient and effective training package for users of DBR-SIS. Taken together, this line of research has yielded some substantive recommendations for training. Raters that received training consisting of a moderate amount of practice and feedback assigned more accurate ratings of behavior (when compared to expert consensus scores) than those raters that received training consisting of only brief familiarization or overview of DBR (Schlientz, Riley-Tillman, Briesch, Walcott, & Chafouleas, 2009; Chafouleas, Kilgus, Riley-Tillman, Jaffery, & Harrison, 2012). In addition, extensive training packages did not tend to significantly improve the accuracy of ratings over a standard training package that included a moderate amount of practice and feedback (LeBel, Kilgus, Briesch, & Chafouleas, 2010; Chafouleas et al., 2012). Finally, even after receiving a standard training package, raters tended to inaccurately assess behaviors occurring at medium rates (i.e. those points between 3 and 8 on the DBR scale; Harrison, Riley-Tillman, & Chafouleas, 2014).

**Development of a web-based module.** Considering the recommendations from the relevant rater training research, Chafouleas and colleagues (2015) developed a web based training module designed to improve rater accuracy. In addition to allowing the possibility of improving the accuracy of DBR ratings, a web-based module is logical because it allows for increased dissemination at a low cost (Gregory & Salmon, 2013). In other words, a web-based

module can be accessed at any time by any individual with an Internet connection, and the cost to host and maintain such a module on a website is relatively lower than the cost of providing in-person trainings.

The web based module, or DBR online training module, is comprised of three distinct parts. Participants initially participate in a brief familiarization of DBR, wherein scale composition and conceptualization as well as the requisite procedures to use DBR are explained. Specifically, this section elaborates operational definitions of each core behavior and orients participants to qualitative scale anchors for the endpoints and midpoint of the scale (i.e. 0 = *never*, 5 = *sometimes*, 10 = *always*).

Upon completion of the brief familiarization, participants receive frame of reference training. In broad terms, frame-of-reference training refers to a set of procedures wherein trainees observe a behavior, and trainers discuss which are the most salient aspects of the observed behavior as well as a rationale for assigning a particular rating (Bernardin & Buckley, 1981). In a study by McIntyre and colleagues (1984), undergraduate students were exposed to rater error training (i.e. a discussion of the common types of rating errors that may occur), frame-of reference training, and no training, and they were assessed on their accuracy of ratings of a videotaped lecture. It was found that participants who received frame-of-reference training evidenced better rating accuracy than those who received rater error training or no training. During frame-of-reference training, participants receive guided practice in which they view video clips of student behavior. After viewing each clip, participants are asked to reflect upon the video and generate a possible rating. Following this brief period of thought, the correct rating, established using an expert consensus building procedure (Jaffery et al., 2015), is displayed to participants along with an explanation of why that particular rating was the correct value.

The final section consists of independent practice and feedback. This section consists of three "main" clips and up to 18 "auxiliary" clips. All participants view each main clip regardless of their rating accuracy. Each main clip asks participants to simultaneously rate academically engaged, respectful, and disruptive behavior. If a participant assigns a rating that is within the allowable range about the correct score (defined as +/- one DBR point for end point ratings or +/- two DBR points for mid-point ratings), then a participant is praised and allowed to proceed to the next main clip. However, if a participant does not assign a rating within the allowable range, then that participant receives up to two auxiliary clips for each behavior that he inaccurately rated. These auxiliary clips ask participants to only rate one behavior at a time. Following both accurate and inaccurate ratings, the module provides feedback, citing specific examples, regarding why the correct rating of the behavior was a particular value. It is important to note that regardless of the amount of auxiliary clips a participant needs, each participant has an opportunity to rate every behavior type occurring at high, medium, and low rates. A branching chart is listed in Appendix B to further clarify the sequence of clips that a participant receives (Chafouleas et al., 2015).

In order to directly test the influence of the training module on DBR-SIS rating accuracy, Chafouleas and colleagues (2015) designed a study that consisted of 90 undergraduates rating video clips of student behavior. Results from this study have provided some preliminary support for the effectiveness of the DBR online training module in improving rater accuracy (Chafouleas et al., 2015). In this study, participants were assigned to either an experimental group (those who viewed the DBR online training module) or a control group (those who participated in a brief familiarization of assessing student behavior using DBR scales). Analyses determined that the ratings of participants in the experimental group corresponded more closely with scores obtained

from DBR experts and SDO than did ratings of participants in the control group. It is important to note, however, that improved rating accuracy was not found for all behavior types, rates, and comparisons. Although results generally favored the module-trained raters across all rating situations, statistically significant improvements in accuracy were only found for ratings of high rates of academic engagement and respectful behavior when using both DBR and SDO as the comparison score. Statistically significant improvements in accuracy were also evident for medium rates of respectful behavior, but only when SDO was used as the comparison score.

Although Chafouleas and colleagues (2015) conducted important step in the validation of online training procedures, it is limited in a number of ways. Primarily, participants based their ratings on the viewing of several video clips. Given the intended purpose of DBR (i.e. assess student behavior as it occurs in the classroom), it would be ideal to assess the influence of any training procedure on the assessment of *in vivo* behaviors. Additionally, although accuracy is a critical property of any behavior rating, it should not be assumed that an improvement in accuracy is akin to an improvement in reliability. However, it is conceptually possible that completion of the DBR online training module would improve reliability of scores as well. As a rater becomes more likely to assign the correct score (more accurate), it follows that they are more likely to assign that score more consistently (more reliable). Thus, the present study seeks to expand upon the original study by providing an opportunity to (a) assess behavior *in vivo* and (b) examine the training module's influence on the reliability of DBR-SIS using a contemporary approach.

**Reliability in behavior assessment**

Traditionally, reliability has been a psychometric topic that has been explored at length in the behavior assessment literature. A critical tenet of any assessment tool is its ability to produce

reliable ratings of behavior. In other words, the instrument should yield consistent results when analyzing the same behavior under the same conditions (Hintze, 2005). The reliability of behavior assessment data has typically been evaluated through the lens of classical test theory (CTT). Within a CTT framework, it is assumed that an observed or measured score is comprised of two distinct components: the true score (i.e. the actual value) plus some amount of measurement error (Brennan, 2011). An instrument that produces reliable data minimizes measurement error and tends to consistently reflect the true score of an individual. However, authors have suggested that the CTT framework may not be adequate for evaluating educational assessment tools (Cronbach et al., 1972). When assessing behavior in applied settings (i.e. school), there is typically a high degree of variability amongst many facets (e.g., rater, day, occasion) that may influence the consistency of any behavior rating (Briesch, Swaminathan, Welsh, & Chafouleas, 2014). When using formative assessment measures, it is logical that there is a certain degree of variability or error that will occur over time, given that ratings are occurring on different days and at different times (i.e. under different conditions; Cone, 1981). Although this type of error variance may even be desirable in formative assessment measures, a behavior assessment tool that is highly susceptible to rater error would limit the capacity of the instrument, as it would make generalization of ratings across raters difficult. Unfortunately, CTT does not allow examination of multiple sources of error variance, but rather groups all sources of error into one single error term (Brennan, 2011).

**Generalizability Theory**

In contrast to CTT, generalizability theory (GT) offers an option to calculate reliability wherein it is possible to simultaneously and individually examine multiple sources of error variance (Cronbach, Rajarnatam, & Gesler, 1963). Unlike in CTT where the goal is to quantify a

true score, GT attempts to estimate a universe score, or an average value of a particular behavior across all conditions (Briesch et al., 2014). The analysis results in a generalizability (G) and dependability (D) coefficient, which reflect how accurate the generalization of a person's observed score is to the universe score and vary depending on the type of decision being made (Shavelson & Webb, 1991).

In utilizing GT, a researcher performs two studies: a generalizability (G) study and a decision (D) study. The purpose of a G study is to estimate the amount of variance attributable to each specific facet (Briesch et al., 2014; Shavelson & Webb, 1991). A facet is any component that can be thought of as contributing substantial variance. Within the realm of behavioral assessment, rater, form, method, and occasion are typically facets of interest. When performing a G study, it is critical to define specific characteristics of each facet being studied. The first critical distinction is to determine whether or not the facets of interest are fully crossed or nested (Briesch et al., 2014). A fully crossed facet is one in which all conditions of one facet co-occur with all conditions of another facet. For example, in a study in which the facets of interest are person, rater, and occasion, these facets would be considered fully crossed if the same raters rated all persons on every occasion. On the other hand, if each rater rated different persons, persons would be considered nested within rater. A second essential distinction is to determine whether facets are random or fixed. Denoting a facet as random implies that the facet is a sample from the universe of observations, while designating a facet as fixed implies that it is completely representative of the entire universe (Brennan, 2011).

After completing the G study, a D study uses the results of the G study to inform the best possible measurement for a specific purpose (Shavelson & Webb, 1991). Among other considerations, a D study can inform the number of data points that are necessary to achieve a

reliable estimate of behavior (Chafouleas et al., 2010). To this extent, the D study results in a generation of both a generalizability (G) coefficient represented by $E\hat{\rho}^2$ and a dependability (D) coefficient represented by $\Phi$. Interpretations of the G and D coefficients, however, differ by the type of decision that is being made (i.e. relative or absolute decisions).

G coefficients are most useful to inform relative decisions or those relied upon for screening purposes. It is possible to examine changes in G coefficients as a function of the number of observations obtained, thereby allowing an analysis of the number of observations required to reach a given reliability threshold. It is important to note that the number of data points required to achieve an acceptable G coefficient (i.e. reach the specified reliability threshold) is the number of data points needed to generate one reliable data point. Thus, for the purposes of screening, if the G coefficient approaches adequate reliability after five observations, then an estimate derived from five observations is needed to reliably rank-order a student (i.e. determine at which percentile rank the student's score would fall).

On the other hand, D coefficients are often interpreted to inform absolute decisions or those relied upon for progress monitoring purposes (Salvia, Ysseldyke, & Bolt, 2010; Briesch et al., 2014). Similar to the G coefficient, it is possible to examine the number of observations required to achieve a desirable D coefficient, which is equivalent to the number of data points needed to generate one reliable data point. In the context of progress monitoring, however, school psychologists are most interested in reliably assessing change across time. It has typically been recommended that five reliable data points are desired to demonstrate change across time. Thus, if the D coefficient approaches adequate reliability after five observations, then twenty-five observations are needed to obtain five reliable data points (averaging over every five observations).

**Investigations of DBR-SIS reliability using Generalizability Theory**

Generalizability theory has been previously utilized to examine the reliability of DBR-SIS. Chafouleas and colleagues (2007) conducted the first examination of DBR-SIS within a GT framework. In this study, raters received a brief familiarization of DBR-SIS instrumentation and use before participation. The raters utilized a DBR-SIS scale to assess two student behaviors: working to resolve conflicts and interacts cooperatively) the researchers found that there was a substantial proportion of variance attributable to the facet of rater (40% of the variance in ratings of works to resolve conflict and 20% of the variance in interacts cooperatively). In addition, results indicated that between four and seven data points were needed to achieve G and D coefficients of .70 (i.e. adequate for low stakes decisions) and 10 data points were needed to achieve G and D coefficients of .90 (i.e. desirable for high stakes decisions). Briesch and colleagues (2010) echoed these results in a similar study wherein it was found that over one quarter of the variance of DBR-SIS ratings of academic engagement were due to rater related effects. Given the divergence in raters' judgments of behavior, results from these studies suggested that DBR-SIS ratings must be analyzed within individual rater over a series of observations, thereby relying on an individual's perception of a problem (Chafouleas et al., 2007; Briesch et al., 2010). Therefore, it has been recommended that DBR-SIS data not be interpreted across multiple raters. The alternative to this interpretation of findings would require ratings over a shorter period of time to be averaged over a larger number of raters. Although this may be possible in some settings, the use of multiple raters to assess the same behavior at the same time defeats the major benefit of using DBR-SIS—its ability to be completed repeatedly without being overly burdensome on school resources.

Other recent GT based evaluations of DBR-SIS have yielded somewhat different results. A study involving middle school students and two different types of raters (i.e. two teachers and two graduate students in school psychology) demonstrated an overall reduction in error variance attributable to rater (Chafouleas, Briesch, Riley-Tillman, Christ, Black, & Kilgus, 2010). In this study, teachers received a brief familiarization of DBR-SIS use and instrumentation, while the graduate students had substantial experience and training in behavior assessment. Results from overall model suggested that rater error was responsible for only 5% of the variance in ratings of academic engagement, and only 2% of the ratings of disruptive behavior. However, there was still some significant between-rater variation in the ratings conducted by the two teachers. Most notably, rater effects accounted for 8% of the variance in teacher ratings of academic engagement, while they accounted for 0% of the variance of researcher ratings of academic engagement. Furthermore, this study indicated that, when averaging across both teacher ratings, 15 observations were required to obtain adequate reliability for relative decision-making ($E\hat{\rho}^2 >$ .80), and that sufficient reliability for absolute decision-making ($\Phi >.70$) could not be achieved in 20 observations. However, when analyzed at the level of the individual teacher, it was found that one teacher could achieve adequate reliability for both relative and absolute decision making in roughly 10 observations, whereas the other teacher required upwards of 20 observations to reached bare minimum levels of reliability for both purposes of decision-making. Given these results, recommendations from this study agreed with earlier evaluations in that interpretation of DBR-SIS data should be conducted within rater. It is important to note that, across outcome measures, the graduate students (i.e. those who had received training in the course of their education) tended to produce more reliable ratings than the teachers. However, teachers in both GT evaluations of DBR-SIS did not receive any type of formal training. Therefore, it is possible

that, with training, teachers could produce data with increased reliability. Thus, further investigation is necessary to examine methods aimed at improving reliability of DBR-SIS data in order to allow (a) flexible use across raters and (b) increased efficiency in decision making.

**Purpose of Study**

DBR-SIS has emerged as a useful assessment method within an MTSS model as it has demonstrated technical adequacy as a screener and a progress monitor. However, DBR-SIS is not without its limitations. Previous findings have suggested that the utility of DBR-SIS is limited in (a) assessing behavioral progress in situations where students attend classes with different teachers (i.e. secondary settings) and (b) situations where it is necessary to quickly obtain a reliable estimate of behavior. In the absence of empirically supported training methods to increase reliability, these limitations on DBR-SIS use will persist. However, it is conceptually possible that completion of the DBR online training module has the capacity to reduce these restrictions on DBR-SIS use. To this extent, the purpose of this study is to expand upon the initial evaluation of the DBR online training module in two critical ways. Primarily, the present study utilizes a generalizability theory framework to evaluate how completion of the training module influences the reliability of DBR-SIS data. In addition, this study utilized certified teachers as the participant raters and allowed them the opportunity to rate *in vivo* behaviors. Research questions and corresponding hypotheses were as follows:

**Research question 1.** Do rater types (module trained v. brief familiarization only) demonstrate differences in proportions of variance attributable to each facet?

***Hypothesis 1.*** Teachers receiving training using the DBR training module will demonstrate greater proportions of variance due to the individual student and occasions of

measurement (and their interactions) whereas teachers that receive a brief familiarization only will demonstrate greater proportions of variance due to rater and error.

**Research question 2.** Do results suggest that individual raters within each rater type use DBR-SIS similarly?

***Hypothesis 2.*** Teachers trained using the DBR training module will demonstrate similar patterns of use of DBR-SIS, allowing for the possibility of generalization across raters. Teachers given a brief familiarization only will demonstrate individual differences in their use of the instrument, consistent with previous GT evaluations of DBR-SIS.

**Research question 3.** Is there a difference in the number of observations required to achieve sufficient levels of reliability for absolute and relative decision making when comparing across rater type and individual rater?

***Hypothesis 3.*** Teachers trained using the DBR training module will be able to achieve sufficient levels of reliability for both absolute and relative decision making in fewer observations than their counterparts that receive a brief familiarization of DBR. Individual teachers that were trained using the online module will show similar levels of reliability after the same number of observations, whereas individual teachers receiving a brief familiarization may show some discrepancies.

## Method

### Participants

Participants in this study included four teachers and six elementary-aged children. All participants were members of a summer day camp located in the Northeast United States. Teacher recruitment occurred during a meeting of camp staff before the summer season began as they were given recruitment flyer inviting them to participate in the study As teachers are the

most likely users of the DBR, all raters in the study were certified teachers. MT rater two and BF rater one were certified in physical education, whereas BF rater one was certified in general education and BF rater two was certified in music. The teachers were White, non-Hispanic males, with the exception of BF Rater two who was female. All teachers had between one and five years of experience and possessed at least a Bachelor's degree, though MT rater two and BF rater two possessed Master's degrees. Potential child participants consisted of all campers that were scheduled to participate in the mid-morning ice skating activity for at least three consecutive weeks. Parents or guardians of approximately 25 children were contacted about the study via a hard mailing before the beginning of the camp season. Parents or guardians were given a recruitment flyer that elaborated on the study procedures and requested parental consent to participate in the study. However, more than the required number of participants (six) indicated their desire to participate in the study. Therefore, the six children that were selected for participation in the study (i.e. those whose behavior be rated) were nominated through a process in which the researcher and the teachers worked together to identify those children most likely to demonstrate variability in their behavior to facilitate data analysis as necessary within the planned design. All children were six or seven-year-old males from White, non-Hispanic backgrounds, with the exception of camper two who was Hispanic. Each child had attended the day camp for either two or three years; therefore they were familiar with the daily routine and expectations.

**Setting**

The study took place at a summer day camp located in the Northeast United States. The day camp runs for eight weeks, and consists of approximately 700 total children and 200 staff members. Participants were observed and rated during their ice-skating activity, which occurred at approximately 10:30 AM each day. The duration of each daily ice-skating session was approximately 30 minutes. Ice-skating was chosen as the target setting due to the inherent potential for variability in participant behavior. Due to the nature of the activity (e.g., enjoyable but physically fatiguing), it was likely that campers will demonstrate engagement in the activity, but it was unlikely they will be engaged for the entire period. The activity was also relatively unstructured (i.e. less supervision when compared with a classroom setting). As such, campers had increased opportunity to demonstrate behaviors that did not comply with posted rules. In other words, there was ample opportunity for campers to demonstrate a wide range of behaviors within a setting where all campers were participating in the same activity.

**Measures**

Although DBR-SIS has traditionally been used as an assessment of classroom behavior, its flexibility allows for its use in multiple settings and across a range of target behaviors. Given that behavior expectations are relatively similar in both day camp and school (i.e. be engaged, do not be disruptive), it is logical that DBR-SIS can be used to assess behavior in this setting. Activity engagement (AE) was a primary measure in this study. Similar to active academic engagement in the classroom setting as defined by Shapiro (2004), activity engagement was defined as actively participating in the ice-skating activity (e.g., skating, putting on equipment). Non-examples of activity engagement included being off the ice for any reason other than to fix equipment as well as being on the ice, but not skating (e.g. standing against the boards). Disruptive behavior (DB) was the other primary measure used in this study. DB was defined as

action that interrupts regular ice skating activity. Examples of disruptive behavior included skating in the wrong direction, playing tag on the ice, throwing snow at other individuals, or any other violation of posted rules. An example of the DBR-SIS form that was used for the purposes of this study can be found in Appendix C.

**Recruitment and selection of participants**

The researcher first sought teacher participation. Prior to the beginning of the camp season, an advertisement was distributed to all of the teachers employed by the summer camp. Recruitment continued until four teachers consented to participate in the research study. Once teacher participation was secured, the researcher contacted all children that were scheduled to participate in the 10:30 AM skating session for at least three consecutive weeks of camp via a hard mailing letter. The letter described the nature of the study and request consent for the child to participate. Upon receipt of the consent forms, a list of participating children was compiled. In order to achieve sufficient data for the planned analyses, it was necessary that each child participant demonstrate some variability in their behavior. Therefore, the primary researcher and participating teachers worked together to nominate campers who were most likely to demonstrate a range of behaviors. This process continued until six campers were identified for further participation. Campers that gave consent, but were not nominated for further participation, were exited from the study.

**Rater training**

Prior to any observations at the summer day camp, all four raters participated in rater training. Raters were randomly assigned to the type of training that they received. Two raters completed the DBR online training module (henceforth referred to as module trained; MT), whereas the other two raters received only a brief familiarization of DBR-SIS use and

instrumentation (henceforth referred to as brief familiarization; BF). As mentioned previously, the DBR online training module consisted of an overview of DBR use, frame of reference training, and opportunities for practice and feedback. The brief familiarization of DBR use and instrumentation included a discussion of the DBR-SIS scale, operational definitions of each behavior, and relevant rating procedures (e.g., complete ratings immediately following the observation period). It is important to note that two teachers were included as members of each rater type in order to allow for the examination of rater-related effects. After the completion of rater training, the primary researcher ensured all raters understood study procedures.

**Data collection**

Data were collected over 10 days. Each rater simultaneously rated the active engagement and disruption of each camper two times per each ice-skating period. Teachers were able to discriminate those campers in the study due to prior familiarity with the helmet each camper wore on the ice. The first rating pertained to the first half of the activity, and the second rating was based on the second half of the activity. Given that the entire activity period was 30 minutes long and raters required 2-3 minutes to complete the ratings, each observation was based on approximately 12-13 minutes of actual time. Raters were stationed in the bleachers that look down on the rink, and they did not leave at any point during the observation period. The bleachers allowed each rater an unobstructed view of all campers at all times. Raters were spaced out along the length of the bleachers as to minimize the opportunity for inter-observer reactivity. Given the number of students that were on the ice (i.e. at least 50) for the ice-skating session, it is unlikely that participating students displayed any substantial amount of reactivity.

Each rating period began upon the commencement of the mid-morning ice-skating session. Raters were given cues to denote when it was time to begin an observation period as

well as when it was time to complete a rating. This was done to ensure that each rater observed each child for the same length of time. The order in which raters assessed each camper was randomized for each observation. After each rating day (i.e. after the ice skating activity ends), the primary researcher collected all ratings and stored them in a locked compartment.

**Design and analysis**

Ratings were analyzed using GT. The design involved facets of person, rater, day, occasion, and behavior. The facets of person, rater, day, and behavior were considered fully crossed (i.e. all raters will rate both behaviors of all students on all days), but the facet of occasion was considered as nested within day since two observations within a particular day were not independent of each other.

Given that each behavior in the study is not necessarily representative of the entire universe of behavior, the researcher employed a model that treated behavior as a fixed facet. Therefore, step one in the analysis consisted of the generation of two identical models for each behavior (i.e. behavior treated as a fixed facet; one model for activity engagement and one model for disruptive). This model echoed that employed by Chafouleas and colleagues (2010). In this particular study, each model for each behavior carried the potential for 480 data points. A visual depiction of the facets and relevant interactions that comprise this model is shown in Appendix D. As neither of the remaining facets is completely representative of the entire universe, each remaining facet (person, rater, and occasion nested within day) was treated as random. The equation for this model is as follows:

$$\sigma^2 (X_{prdo}) = \sigma^2_p + \sigma^2_r + \sigma^2_d + \sigma^2_{o,do} + \sigma^2_{pr} + \sigma^2_{pd} + \sigma^2_{rd} + \sigma^2_{pdr} + \sigma^2_{po,\ ro,\ prod,\ e}$$

It is important to note that the model in step one implies that each member of each facet is representative of the universe, but this is only true if one assumes that all raters in this study

are interchangeable with each other. Recall this study aims to uncover differences between raters based on the type of training that each receives. Therefore, step two of the analysis involved exploration of differences between rater types (MT vs. BF). Thus, this analysis consisted of a descriptive comparison of four models (AE x MT, AE x BF, DB x MT, DB x BF) that were comprised of the random facets of person, rater, and occasions of measurement. Each model at this step in the analysis allowed for a maximum of 240 observed data points. A visual depiction of this model is displayed in Appendix E. The equation for this model is identical to the model in step one:

$$\sigma^2 (X_{prdo}) = \sigma^2_p + \sigma^2_r + \sigma^2_d + \sigma^2_{o,do} + \sigma^2_{pr} + \sigma^2_{pd} + \sigma^2_{rd} + \sigma^2_{pdr} + \sigma^2_{po,\ ro,\ prod,\ e}$$

Through a set of G studies completed for each of the four models in step two, it was possible to analyze descriptive differences in the proportions of variance attributable to each facet for each behavior and rater type (research question one). Upon analysis of the variance components, special attention was paid to the proportion of variance attributable to person ($\sigma^2_p$), day ($\sigma^2_d$), and rater ($\sigma^2_r$). An ideal behavior assessment tool should maximize the percentage of variance attributed to $\sigma^2_p$ and $\sigma^2_d$, as this would indicate that persons systematically varied in their behavior across observations—an expected result. On the other hand, an ideal training method should minimize the value of $\sigma^2_r$. One can estimate the relative value of $\sigma^2_r$ by computing the square root of $\sigma^2_r$ in order to estimate the expected range of rater means on a given scale (Shavelson & Webb, 1991). Under the most ideal measurement conditions, there would be a range of zero. This would indicate that raters always used the same part of the scale to rate the same behavior under the same conditions.

Following this group of G studies, a series of D studies was conducted using the models introduced in step two. This set of D studies allowed the generation of both $E\hat{\rho}^2$ and $\Phi$

coefficients. As mentioned previously, $E\hat{\rho}^2$ is most typically used to inform decisions about screening, whereas Φ is most often used to inform decisions about progress monitoring. For the purposes of the analysis at step two, $E\hat{\rho}^2$ was computed as:

$$E\hat{\rho}^2: \frac{\sigma^2 p}{(\sigma^2 p + \sigma^2 Rel)}$$

$$\text{Where } \sigma^2_{Rel} = \frac{\sigma^2 pr}{n_r} + \frac{\sigma^2 pd}{n_d} + \frac{\sigma^2 pdr}{n_d n_r} + \frac{\sigma^2 po,ro,prod,e}{n_r n_d n_o}$$

And Φ was computed as:

$$\Phi: \frac{\sigma^2 p}{(\sigma^2 p + \sigma^2 Abs)}$$

$$\text{Where: } \sigma^2_{Abs} = \frac{\sigma^2 r}{n_r} + \frac{\sigma^2 d}{n_d} + \frac{\sigma^2 o,do}{n_d n_o} + \frac{\sigma^2 pd}{n_d} + \frac{\sigma^2 pr}{n_r} + \frac{\sigma^2 rd}{n_r n_d} + \frac{\sigma^2 pdr}{n_d n_r} + \frac{\sigma^2 po,ro,prod,e}{n_r n_d n_o}$$

Descriptive analysis took place between rater types, focusing specifically on the number of observations required to achieve adequate levels of reliability for both absolute and relative decision making (research question three). According to Salvia, Ysseldyke, & Bolt (2010), the minimum threshold sufficient for absolute decision-making (progress monitoring) is .70, while .80 is recommended for relative decision-making (screening). Similar to the strategy employed by Chafouleas et al. (2010), $E\hat{\rho}^2$ and Φ coefficients are presented in a table and graph across specific intervals of observation (e.g., 1 day, 2 days, 5 days, 10 days), thus enabling comparison of coefficients between rater types (i.e. aggregated across both raters in each type) at selected intervals.

After exhausting the analyses at step two, step three involved descriptive analysis of the model reduced by individual rater. This final step was critical as it more closely represents the conditions under which DBR-SIS data are collected (i.e. in the classroom by a single teacher; Chafouleas et al., 2007). Thus, step three involved the comparison of eight models (i.e. AE x MT rater one, AE x MT rater two, AE x BF rater one, AE x BF rater two, DB x MT rater one, DB x

MT rater two, DB x BF rater one, DB x BF rater two) which were comprised of the random facets of person and occasions of measurement. Each model at this step allowed for a maximum of 120 data points. A visual depiction of this model is shown in Appendix F. The equation for each of the step three models is as follows:

$$\sigma^2 (X_{pdo}) = \sigma^2_p + \sigma^2_d + \sigma^2_{o,do} + \sigma^2_{pd} + \sigma^2_{po,\,e}$$

After the generation of these eight models, another series of G studies was conducted. The analysis of the G studies conducted at step three consisted of analyzing descriptive differences in the proportions of variance attributable to each random facet (i.e. person, day, occasion: day, and relevant interactions) for each behavior type and individual rater. It was desirable that the proportion of variance attributable to each random facet be similar for each rater within the MT group. If both raters evidenced a similar pattern of variance distribution, it would suggest that the individual raters are using DBR-SIS in a similar manner (research question two).

Using the same models generated at step three, an additional set of D studies was conducted to obtain a generalizability coefficient ($E\hat{\rho}^2$) and a dependability coefficient ($\Phi$) for each individual rater. At step three, $E\hat{\rho}^2$ was computed as:

$$E\hat{\rho}^2: \frac{\sigma^2 p}{(\sigma^2 p + \sigma^2 Rel)}$$

$$\text{Where } \sigma^2_{Rel} = \frac{\sigma^2 pd}{n_d} + \frac{\sigma^2 po,e}{n_d n_o}$$

And $\Phi$ was computed as:

$$\Phi: \frac{\sigma^2 p}{(\sigma^2 p + \sigma^2 Abs)}$$

$$\text{Where: } \sigma^2_{Abs} = \frac{\sigma^2 d}{n_d} + \frac{\sigma^2 o,do}{n_d n_o} + \frac{\sigma^2 pd}{n_d} + \frac{\sigma^2 po,e}{n_d n_o}$$

Similar to the method described above (and utilized in Chafouleas et al., 2010), $E\hat{\rho}^2$ and $\Phi$ are displayed in a table and graph across specific intervals of observations. However, these descriptive comparisons of both reliability-like coefficients occurred at the level of each individual rater. Thus, this analysis described the amount of observations an individual rater required to achieve a given level of reliability for relative and absolute decision-making (research question three). It was desirable that both raters within the module-trained group will display sufficient levels of reliability within similar numbers of observations (e.g., both raters achieve a $\Phi$ coefficient of .70 after twenty observations), thereby suggesting raters were using DBR-SIS in a uniform manner.

## Results

Overall, 906 of 960 data points (94%) were collected. The few instances of missing data were due to camper absence or removal from the ice (e.g. injury) during the ice skating period. Data were entered into SPSS, which is a statistical software package that has the ability to compute variance components (i.e. the G study) on data sets with missing data (Briesch et al., 2014). Given the relative completeness of the dataset and the ability of SPSS to handle missing data, it was decided that multiple imputation techniques would not be employed to fill in missing values (Briesch et al., 2014). D studies were computed in Microsoft Excel using the variance component estimates and the formulas listed above.

Overall mean rating patterns are shown in Figure 1. Although some raters were consistent with others for some students, there were typically mean discrepancies in how each rater viewed each student's behavior. For example, MT raters differed in their mean estimates of AE by at least one full DBR point on four out of the six students. It is also interesting to note that, with a few exceptions, some general rank order patterns emerged. MT rater one tended to rate students

on the higher end of the AE scale, while BF rater 1 tended to rate students on the lower end of that same scale. Raters tended to produce somewhat more consistent mean estimates of DB, as MT raters differed by at least one full DBR point on only two out of the six students. However, BF rater one was discrepant from the other raters, as he always perceived the students to have displayed higher levels of DB.

**G Study: Full Scale Model**

The VarComps procedure with Type III Sum of Squares was originally used as an estimation technique. This analysis, however, revealed a number of negative variance components, which were indicative of sampling error. In order to address these negative variance components, the estimation method was changed to Restricted Maximum Likelihood (REML). REML is an estimation method that does not allow for negative variance components and has often been suggested as an alternative procedure in G studies in which there is considerable sampling error (Briesch et al., 2014). Variance components estimates with REML were compared to those made with Type III Sum of Squares to inspect for any substantial differences that may indicate some bias produced by the change in estimation method. However, no such differences were noted. Therefore, REML estimates were accepted for analysis.

The results for the full scale G model which considers all raters as members of the same universe are shown in Table 1. Similar results were identified for both AE and DB, and they show that raters had a difficult time generating a consistent estimate of each behavior. Unexplained error accounted for 47% and 31% of the total variance in AE and DB, respectively. In addition, combinations across components that are ideally maximized in a G study (i.e. person, day, and their interaction) did not account for more than 11% of the total variance for either behavior. The full-scale model also shows large proportions of variance due to the facet of

rater and its interactions. This result is not unexpected, however, given that the purpose of the study was to compare differences in ratings from two different groups of raters. Thus, it was expected that raters would have different perceptions of behavior based on their respective training.

**Reduced Model by Rater Type: G Study**

As mentioned previously, random facets in full model G study assume that each specific person selected for participation is representative of the universe that one wishes to generalize. Given that raters were trained differently and the main goals of this study (i.e. the research questions) are to draw conclusions about a specific type of rater, it was necessary to conduct a G study for each rater type.

The results from the reduced model by rater type are presented in Table 2. The model identified some similarities and differences between rater types for AE. It was originally proposed that MT raters would evidence more variance due to person and day (and their interactions), while BF raters would evidence more variance attributable to rater and error. Module trained raters did evidence slightly more variance for the components of person, day, occasion: day, and person x day, but the magnitude of these differences was not particularly pronounced (MT = 10%; BF = 9%). Thus, module trained raters were only slightly better in discriminating behavioral differences among students and using the AE scale consistently across time. Conversely, the model identified substantially more variance attributable to rater for BF raters (48%) than MT raters (17%), meaning that MT raters tended to be more consistent with each other than BF raters when rating AE. However, the model also identified more unexplained error variance for MT raters (50%) than BF raters (38%).

Similar patterns between rater types were noted for DB. MT raters were better able to discriminate behavioral changes and tended to use the scale more consistently across time (29% of total variance across day, occasion: day, and person x day components) than BF raters (5% of total variance across the same components). BF raters evidenced substantially greater inconsistency in overall ratings of DB (67% due to rater component) than MT raters (0%). However, MT raters did again show more unexplained error variance (60%) than BF raters (16%). Finally, neither rater type was able to discriminate between campers' levels of DB, as the model identified no variance (0%) attributable to person.

**Reduced Model by Individual Rater: G Study**

The results of reduced model by individual rater are shown in Table 3. This model was explored in order to better understand individual rater similarities and differences in their assessment of AE and DB. Results show that the proportions of variance attributable to each facet were slightly more similar between MT raters for both AE and DB. The proportion of variance attributable to each facet for AE was generally comparable across both MT raters. In addition, MT raters demonstrated greater similarity in the amount of variance attributable to person (MT rater 1 = 11%, MT rater 2 = 5%), occasion: day (0% for both), and unexplained variance (MT rater 1 = 51%, MT rater 2 = 68%) than BF raters. However, BF raters were more similar to each other in the amount of variance attributable to day (0% for both) and person x day (BF rater 1= 11%, BF rater 2 = 12%).

Overall, there were more discrepancies between the raters in each type when examining DB. However, there was still greater similarity between MT raters compared to BF raters. MT raters evidenced more similar distributions of variance for person (both raters = 5%), day (MT rater 1 = 17%, MT rater 2 = 4%), occasion: day (both raters = 0%), and the overall amount of

unexplained variance (MT rater 1 = 78%, MT rater 2 = 55%). BF raters were only more similar to each other on one facet—person x day (BF rater 1 = 25%, BF rater 2 = 1%).

**Reduced Model by Rater Type: D Study**

Results of decision (D) studies for each rater type are shown in Figure 4. These results revealed only one notable difference between reliability like coefficients for the purposes of relative and absolute decision-making. Module trained raters assessing AE displayed higher levels of reliability (.09) than brief familiarization raters assessing the same construct (.01). No differences in reliability were noted after 20 observations for each rater group assessing DB. It is important to note that the overall level of reliability achieved by each group was poor, indicating discrepancies between the raters in each group in their attempt to generate a consistent estimate of behavior.

**Reduced Model by Individual Rater: D Studies**

Within rater D studies were conducted to analyze generalizability and dependability estimates for each individual rater. The results are shown in Figures 5 and 6 for AE and DB, respectively. Overall, the rater that achieved the highest level of reliability tended to vary by the behavior and the coefficient being examined. After 20 observations, BF rater one generated the most reliable estimate of AE for both relative ($E\hat{\rho}^2 = .68$) and absolute ($\Phi = .67$) decision-making. However, MT rater one demonstrated the most reliable estimate of DB for relative decision making ($E\hat{\rho}^2 = .54$), while BF rater two obtained the most reliable estimate of DB for absolute decision making ($\Phi = .50$). It should also be noted that MT raters tended to show more similar levels of reliability than BF raters at given observation interval when rating activity engagement.

## Discussion

The purpose of this study was to evaluate the influence of online training on the reliability of DBR-SIS scores. Overall, trends in the data suggested that completion of the online training module tends to lead to an improvement in the reliability of DBR-SIS data. Although the present results do not overwhelming support that completion of the DBR online training module results in improved reliability, module-trained raters tended to generate more reliable scores than raters provided only with brief familiarization. Specifically, module-trained raters tended to more reliably assess the range of target behaviors than brief familiarization raters. However, no firm conclusions about the effectiveness of the training module can be drawn due to the raters' difficulty discriminating behavioral differences among students and meeting desired thresholds of reliability. Thus, although results do not indicate that completion of the training module fully remediates rater error, trends support that completion of the training module is related to an improvement in the reliability of DBR-SIS data.

The first research question examined whether rater types would demonstrate differences in the proportions of variance attributable to each facet. The reduced model G study by rater type lends some support to the first hypothesis, as MT raters tended to display slightly greater proportions of variance due to person, day, and their interactions, whereas BF raters displayed substantially greater proportions of variance to due rater. Thus, even though MT raters demonstrated overall low proportions of variance due to the facet of persons, day, and their interactions, relative to the facets of rater and error for both AE and DB, they displayed a more favorable distribution of variance than BF raters. However, it is still somewhat difficult to draw firm conclusions given (a) the magnitude of differences and (b) the finding that MT raters demonstrated a greater proportion of variance due to unexplained error for both behaviors.

The second research question explored whether raters within each rater type would use DBR-SIS similarly. In an ideal setting wherein raters were truly interchangeable members of the same universe, each individual rater should demonstrate similar distributions of variance across relevant facets. Similar to the reduced model by rater type, the results from the reduced model by individual rater showed slight support for the superiority of the MT raters despite an undesirable distribution of variance (i.e. facets of person, day and interactions were small relative to unexplained error). MT raters showed greater similarity in their estimated proportions of variance across the majority of facets on both behaviors. However, this finding is tempered by the overall magnitude of the similarities as well as the instances in which BF raters evidenced more uniform variance distribution. Thus, although trends support hypothesis two, it is difficult to make concrete statements about the extent to which completion of the online training module influences raters to assess behavior in a similar manner.

Finally, the third research question investigated the number of observations required for each rater type and individual rater to generate a reliable estimate of behavior for both absolute and relative decision making. Both rater types and each individual rater struggled to reach adequate levels of reliability for both absolute and relative decision-making given that they all demonstrated low proportions of variance attributable to person. This finding was unexpected given that these thresholds were met in previous GT evaluations of DBR-SIS, even in cases where a substantial amount of rater error was found (Chafouleas et al., 2007; Chafouleas et al., 2010). However, the data still lend some support to hypothesis three. Even though MT raters did not achieve higher reliability when assessing DB nor did MT raters always achieve the highest level of individual reliability, data on activity engagement from MT raters as a group demonstrated greater levels of reliability than those produced by BF raters.

Finally, given the similarity between this study and Chafouleas et al. (2010), it is important to compare results. Both studies suggested that raters with formal training tended to produce more reliable ratings relative to those without any such training. In addition, both studies found that teachers without any formal training tended to diverge in their perception of student behavior. In examining mean ratings of students by each rater, both the present study and Chafouleas et al. (2010) show evidence of at least one rater in the brief familiarization condition who seemed to anchor ratings at a distinctly different point of the DBR scale. Results of the present study show that BF rater 1 always tended to perceive students as being more disruptive and less engaged than the other raters, whereas Chafouleas and colleagues (2010) found that a consultant teacher to anchor ratings substantially higher on the academically engaged scale. Thus, although these differences in anchoring cannot be directly attributed to each teacher's training (or lack thereof), it is worth noting that similar patterns were seen across two different studies for teachers that received only a brief familiarization.

**Limitations**

A few possible reasons might exist as to why MT raters had difficulty discriminating behavioral differences among students and reaching adequate levels of reliability, and as to why they did not do so consistently and substantially better than BF raters. First, it is possible that completion of the DBR online training module only has the capability to foster slight improvements in reliability, but it does not possess the capacity to alleviate a sufficient amount of rating error that is associated with DBR-SIS. To this extent, the results showed an extremely small amount of variance attributable to the object of measurement (person) in all G studies, which caused reliability estimates to be low. Given this result, it is difficult to make firm interpretations of the other parameter estimates in the study, rendering it difficult to draw clear

conclusions from the data. However, this finding is unique in the DBR-SIS literature, as previous studies conducted with classroom behavior have shown substantial amounts of variance attributable to the object of measurement and substantially higher levels of reliability (Chafouleas et al., 2007, Chafouleas et al., 2010). Thus, it is important to consider other factors that may have decreased the amount of variance attributable to the object of measurement and overall reliability estimates. Therefore, it is necessary to engage in further discussion of why these results may have occurred in this particular case, rather than to dismiss the utility of the DBR online training module.

One potential explanation for the results is the overall amount of variability in behavior displayed by each student. Students tended to exhibit a wide range of rates of AE and DB across observation sessions (shown in Figures 2 and 3, respectively). In addition, raters frequently assessed behaviors as occurring at medium rates, whereas end point ratings (i.e. 0 or 1, 9 or 10) occurred somewhat less frequently. Thus, it can be said that raters had greater difficulty producing reliable data for behaviors that required them to use the midpoints of the DBR scale. This result is unsurprising given findings from previous generalizability studies of DBR-SIS (Chafouleas et al., 2010) in which raters also tended to struggle with midpoint ratings, whereas they tended to evidence greater consistency with each other when rating behaviors near the end points of the DBR scale. Taken together, the number of ratings that required raters to use the midpoints of the DBR scale in this particular sample of the campers' behavior may have attenuated reliability estimates.

Given the amount of behavior that required raters to use the midpoints of the DBR scale, it is also important to consider the rationale for this study as well as previous evaluations of the DBR online training module. It was argued that an improvement in reliability of scores would be

fostered through an improvement in rating accuracy. To that extent, Chafouleas and colleagues (2015) found that the DBR online training module was most effective in improving accuracy for *high* rates of academic engagement (i.e. between 8 and 10 on the DBR scale). Thus, the most robust improvements in reliability should have theoretically occurred when assessing AE at a high rate. However, the descriptive statistics provided indicate that the rate of AE for each camper across many observations occurred at a *medium* rate (i.e. between 3 and 7 on the DBR scale). Taken together, given less evidence for improved accuracy for medium rates of AE, it follows that completion of the training module would result in relatively smaller improvements in reliability when assessing that particular behavior.

An additional factor to examine regards the practicality of observing the behavior of six children during an active activity such as ice-skating. Although previous evaluations of DBR-SIS have shown raters to be generally capable of adequately rating six students simultaneously, this assumption may not necessarily hold true when those children are participating in a sport. In a classroom setting, each child is generally expected to sit in his seat, and, if there is movement (e.g. disruptive behavior), it is restricted to the confines of the classroom. Thus, a teacher is able to easily locate a particular student to observe their behavior due to previous expectation and knowledge of where that student is located. In addition, if two teachers are engaging in whole group instruction, it is likely that they will both observe most of the target behaviors displayed by the target students. However, these expectations proved not necessarily tenable during an activity like ice-skating. For example, even though the raters were only focused on the observation of the target students during the rating periods (instead of also focusing on teaching a lesson), they reported that they needed to spend a few seconds to locate each target student periodically throughout duration of the observation. In addition, the size of the setting made it

difficult for raters to always simultaneously observe multiple students. In other words, during each observation period, raters often spent a few moments observing an individual student, before transferring their attention to a different target student and observing that student. As this process continued across each observation, it is possible that some raters observed a target behavior displayed by a particular camper, while other raters had their attention focused on a different camper. Thus, the possibility exists in which raters may have (1) based their ratings on an unrepresentative sample of behavior as they did not see every target behavior during the rating period and (2) based their ratings on target behaviors different from those observed by other raters. In other words, raters may have demonstrated a form of halo error, in which their ratings were based on a general impression of a student based on a small sample of behavior (Feely, 2002).

Finally, as all raters were certified teachers, caution should be used in interpreting the study results as relevant for raters that may not have teacher certification (e.g., paraprofessionals). In addition, trends to support MT raters may not be present if the rater's primary task is instruction, rather than observation of a target student (as may be the case in practice). Furthermore, it is important to note that all camper participants in this study were males, and it is possible that females would have displayed a different range of behaviors. This study is also limited in its ability to generalize these results to other behaviors. Thus, limited conclusions can be drawn about the effects of the training module on the reliability of ratings of other behavior constructs, such as respectful.

**Future Directions**

Although results showed some promise for the potential of the DBR online training module to improve reliability of data from DBR-SIS, substantive findings to alter current

recommendations for DBR-SIS use in practice were not found. Thus, data from DBR-SIS should continue to be interpreted within rater. In addition, users of DBR-SIS should still participate in online training, given known improvements in accuracy and the potential for improvement in reliability. Overall, further research is needed to specify additional recommendations for training to improve the reliability of data generated from DBR-SIS.

However, given the promise of the current results, the potential for online training to improve the reliability of DBR-SIS should not be dismissed. Rather, it is important to undertake further investigation to develop a better understanding of how online training affects reliability and how DBR-SIS functions under different conditions. Three different avenues of research may provide suggestions to the issues raised by this study: (a) exploration of online training to improve the reliability of DBR-SIS used to assess classroom behavior (b) investigation of how to best train raters to improve the assessment of behaviors occurring at medium rates, and (c) a line of inquiry to understand the conditions necessary to optimize DBR-SIS use outside of the classroom.

Given the nature of this study, one should not necessarily generalize the results to draw conclusions about the influence of the training module of the reliability of DBR-SIS data generated in a traditional classroom setting. A compendium of evidence has been collected to demonstrate that DBR-SIS is valid for the assessment of classroom behaviors. Thus, several factors related to the ice-skating setting may have deflated reliability estimates in the present study (i.e., a large amount of variability in behavior, differences in expectations between the classroom and ice skating). Therefore, it is prudent to further investigate the capacity of the DBR online training module to improve reliability of DBR-SIS data though studies that focus on the rating of behaviors demonstrated in the classroom setting.

This study is unique in that it is the first investigation of DBR-SIS that required analysis of the midpoint ratings at a greater rate than end point ratings. Thus, although results are surprising in that they diverge from previous literature, they also indicate that the training required for raters to accurately and reliably assess midpoint ratings may be different from the training required to assess end point ratings. To this extent, prior research has not fully developed an understanding of how to best train raters to assess medium rates of behavior. Modifications, such as an increase in the amount of opportunities for practice and feedback using the middle of the scale, may need to be made to the DBR online training module in order to accommodate the challenges that raters face in assessing medium rates of behavior. Empirical evaluations of such modifications meant to amplify the influence of the training module on medium rates of behavior provide an important avenue for future research.

It is important to recognize that the present study is an exploratory investigation of a method to improve DBR-SIS ratings. Beyond being the first attempt to analyze the influence of a training method on the reliability of DBR-SIS, this study represents a significant departure from previous work with DBR-SIS. To this extent, this study represents the first attempt to analyze DBR-SIS ratings generated in a non-classroom setting. Previous work has shown that the validity of DBR-SIS is reliant upon the specifics of DBR instrumentation (i.e. wording of definitions, scaling, behavioral specificity; Chafouleas et al., 2009; Riley-Tillman et al., 2009), so it is logical to assume that the setting or context of the assessment would be an important factor to consider when using the measure. The present results support that assumption. They suggest that it is important to pay attention to the details regarding recommended DBR-SIS use, as those conditions may be necessary for valid assessment. Taken together, the results imply that it may

not be suitable to extend the use of DBR-SIS beyond the specific ways that it has been shown to work in the literature.

In the same way, it is important to consider the capacity of the DBR online training module to be effective for non-classroom behaviors. Although the DBR online training module focuses on behaviors as they are displayed in an academic context, it was assumed that the training would be sufficient to assess ice-skating behaviors as both activity engagement and disruptive were defined similarly to the classroom description and are considered general outcome measures. In other words, engagement is participation in an activity, while disruptive is interrupting a task—regardless of setting. Given the present results, however, it may be important to re-consider that assumption. It is possible that engagement and disruptive are better thought of as a higher order construct, whereas display in a specific activity—academics, ice skating, etc.—are better considered separate factors of those constructs. Thus, in order to produce the most valid measurement possible, separate trainings may need to be geared toward each factor (i.e. the display of the construct in each setting).

Taken together, valid data is required to make accurate decisions regarding student supports. Thus, it is critical to identify methods that allow educators to make sound decisions about student supports as quickly as possible. However, the results of the present study indicate that researchers and practitioners alike must be cautious in the ways they choose to use a particular assessment method, especially when choosing to make slight adjustments to the specifics of an instrument in order to fit a particular situation or need. Although several logical reasons exist to anticipate that DBR-SIS and the DBR online training module would be an effective method to reliably and validly assess student behavior in a non-classroom setting, the present results indicate that the premise that an assessment method will work for any situation at

any time is not tenable. Thus, in the absence of empirical evidence to support an instrument's use in a given situation, one should not simply assume that the data are psychometrically sound. Therefore, the validity of behavior assessment data is contingent upon a measure's use in the appropriate context, and careful attention should be paid to the requisite conditions needed to generate defensible data.

References

Bernardin, H. J , & Buckley, M R (1981) Strategies in rater training. *Academy of Management Review*, 6, 205-212. doi:10.5465/AMR.1981.4287782

Briesch, A.M., Chafouleas, S.M., & Riley-Tillman, T.C. (2010). Generalizability and dependability of behavior assessment methods to estimate academic engagement: Comparison of Systematic Direct Observation and Direct Behavior Rating. *School Psychology Review, 39*(3), 408-421.

Briesch, A.M., Swaminathan, H., Welsh, M., & Chafouleas, S.M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52,* 13-35. doi: 10.1016/j.jsp.2013.11.008.

Brennan, R.L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24,* 1-21. doi:10.1080/08957347.2011.532417

Chafouleas, S.M. (2011). Direct Behavior Rating: A review of the issues and research in its development. *Education and Treatment of Children, 34*(4), 575-591. doi: 10.1353/etc.2011.0034

Chafouleas, S.M., Briesch, A.M., Riley-Tillman, T.C., Christ, T.J., Black, A.C., & Kilgus, S.P. (2010). An investigation of the generalizability and dependability of Direct Behavior Rating Single Item Scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. *Journal of School Psychology, 48,* 219-246. doi:10.1016/j.jsp.2010.02.001

Chafouleas, S.M., Christ, T.K., Riley-Tillman, T.C., Briesch, A.M., & Chanese, J.A.M. (2007). Generalizability and dependability of Direct Behavior Rating to assess social behavior of preschoolers. *School Psychology Review, 36*(1), 63-79.

Chafouleas, S.M., Kilgus, S.P., Jaffery, R., Riley-Tillman, T.C., & Christ, T.J. (2013). Direct Behavior Rating as a school-based behavior screener for elementary and middle grades. *Journal of School Psychology, 51,* 367-385. doi:10.1016/j.jsp.2013.04.002

Chafouleas, S. M., Kilgus, S. P., Riley-Tillman, T. C., Jaffery, R., & Harrison, S. (2012). Preliminary evaluation of various training components on accuracy of Direct Behavior Ratings. *Journal of School Psychology, 50,* 317-334. doi:10.1016/j.jsp.2011.11.007.

Chafouleas, S.M., Riley-Tillman, T.C., & Christ, T.J. (2009). Direct Behavior Rating (DBR): An emerging method for assessing social behavior within a tiered intervention system. *Assessment for Effective Intervention, 34*(4), 195-200. doi:10.1177/1534508409340391

Chafouleas, S. M., Riley-Tillman, T. C., Jaffery, R., Miller, F. G., & Harrison, S. E. (2015). Preliminary investigation of the impact of a web-based module on Direct Behavior Rating accuracy. *School Mental Health, 7,* 92-104. doi: 10.1007/s12310-014-9130-z

Chafouleas, S. M., Riley-Tillman, T. C., & McDougal, J. L. (2002). Good, bad, or in-between: How does the daily behavior report card rate? *Psychology in the Schools, 39,* 157–169. doi:10.1002/pits.10027

Chafouleas, S.M., Sanetti, L.M.H., Kilgus, S.P., Maggin, D.M. (2012). Evaluating sensitivity to behavioral change using Direct Behavior Rating Single-Item Scales. *Exceptional Children, 78*(4), 491-505.

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34,* 201–213. doi: 10.1177/1534508409340390

Christ, T.J., Riley Tillman, T.C., Chafouleas, S.M., & Boice, C.H. (2010). Direct Behavior Rating (DBR): Generalizability and dependability across raters and observations. *Educational and Psychological Measurement, 70*(5), 825-843. doi:10.1177/0013164410366695

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Cronbach, L. J., Rajaratnam, N., & Gelser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology, 16,* 137–163.

Cone, J. D. (1981). Psychometric considerations. In M. Hersen & A.S. Bellack (Eds.*), Handbook of Behavioral Assessment* (pp. 38−68)., 2nd Ed Elmsford, NY: Pergamon Press.

Every Student Succeeds Act of 2015, 20, U.S.C. § 1114 (2016).

Feeley, T.H. (2002). Comment on halo effects in rating and evaluation research. *Human Communication Research, 28,* 578-586.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45,* 117–135. DOI: 10.1016/j.jsp.2006.05.005

Gregory, J. & Salmon, G. (2013). Professional development for online university teaching. *Distance Education, 34*(3), 256-270. doi: 10.1080/01587919.2013.835771

Gresham, F.M. (2011). Response to intervention: Conceptual foundations and evidence based practices. In M.A. Bray and T.J. Kehle (Eds.), *Oxford Handbook of School Psychology*

(pp.), New York, NY: Oxford University Press. doi:10.1093/oxfordhb/9780195369809.013.0185

Harrison, S. E., Riley-Tillman, T. C., & Chafouleas, S. M. (2014). Practice with feedback and base rates of target behavior: Implications for rater accuracy using Direct Behavior Ratings. *Canadian Journal of School Psychology*, *29*(1), 3-20. doi:10.1177/0829573513515424

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review*, *34*, 507-519.

Jaffery, R., Johnson, A.H., Bowler, M.C., Riley-Tillman, T.C., Chafouleas, S.M., & Harrison, S.E. (2015). Using consensus building procedures with expert raters to establish comparison scores of behavior for Direct Behavior Rating. *Assessment for Effective Intervention, 40*(4), 195-204.

Kelley, M. L. (1990). *School-Home Notes: Promoting Children's Classroom Success*. New York: Guilford.

Kilgus, S.P., Chafouleas, S.M., Riley-Tillman, T.C., & Welsh, M.E. (2012). Direct Behavior Rating scales as screeners: A preliminary investigation of diagnostic accuracy in elementary school. *School Psychology Quaterly, 27*(1), 41-50. doi: 10.1037/a0027150.

Kilgus, S.P., Riley-Tillman, T.C., Chafouleas, S.M., & Welsh, M.E. (2014). Direct Behavior Rating as a school-based universal screener: Replication across sites. *Journal of School Psychology, 52,* 63-82. doi:10.1016/j.jsp.2013.11.002

LeBel, T. J., Kilgus, S. P., Briesch, A. M., & Chafouleas, S. M. (2010). The impact of training on the accuracy of teacher-completed Direct Behavior Ratings (DBRs). *Journal of Positive Behavioral Interventions*, *12*, 55-63. doi: 10.1177/1098300708325265

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*, 147-156. doi: 10.1037/0021-9010.69.1.147

Merikangas, K.R., He, J.P., Burstein, M., Swanson, S.A., Avenevoli, S., Cui, L., & Swendsen, J. (2010). Lifetime prevalence of mental disorders in U.S. adolescents: Results from the National Comorbidity Survey—Adolescent Supplement (NCS-A). *Jounral of the American Academy of Child & Adolescent Psychiatry, 49,* 980-989. doi:10.1016/j.jaac.2010.05.017

Riley-Tillman, T.C., Chafouleas, S.M., Christ, T.J., Briesch, A.M., & LeBel, T.J. (2009). The impact of wording and behavioral specificity on the accuracy of Direct Behavior Ratings (DBRs). *School Psychology Quarterly, 24*,1-12. doi: 10.1037/a0015248

Riley-Tillman, T.C., Chafouleas, S.M., Sassu, K.A., Chanese, J.A.M., & Glazer, A.D. (2008). Examining the agreement of Direct Behavior Rating and Systematic Direct Observation data for on-task and disruptive behavior. *Journal of Positive Behavior Interventions, 10*(2), 136-143. doi: 10.1177/1098300707312542

Riley-Tillman, T. C., Methe, S. A., & Weegar, K. (2009). Examining the use of direct behavior rating methodology on classwide formative assessment: A case study. *Assessment for Effective Intervention, 34*, 242-250. doi:10.1177/1534508409333879

Salvia, J., Ysselydke, J. E., & Bolt, S. (2010). *Assessment in Special and Inclusive Education* (11th ed.) Boston, MA: Houghton Mifflin.

Schlientz, M. D., Riley-Tillman, T. C., Briesch, A. M., Walcott, C. M., & Chafouleas, S. M. (2009). The impact of training on the accuracy of Direct Behavior Ratings (DBRs). *School Psychology Quarterly, 24*, 73–83. doi: 10.1037/a0016255

Scholastic. (2012). *Primary Sources: 2012. America's Teachers on the Teaching Profession.* Scholastic, Inc: New York.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Stamoulis, D. T., & Hauenstein, N. M. A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for rate differentiation. *Journal of Applied Psychology, 76*, 994–1003.

Tilly, W. D., III (2008). The evolution of school psychology to science-based practice: Problem solving and the three-tiered model. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology V* (pp. 17–36). Bethesda, MD: National Association of School Psychologists.

Volpe, R.J. & Briesch, A.M. (2012). Generalizability and dependability of single-item and multiple-item direct behavior rating scales for engagement and disruptive behavior. *School Psychology Review, 41*(3), 246-261.

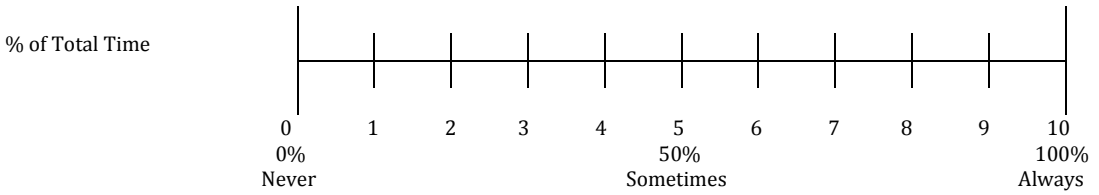
# Appendix A

## Standard Direct Behavior Rating (DBR) Form

| **Date**:<br>_____<br>M T W Th F | **Student Name**: | **Any changes in the typical classroom routine?** If YES, describe (e.g., fire drill, assembly, field trip): |
|---|---|---|
| **Observation Time**:<br>Start:_____<br>_<br>End: _____<br>☐ Check if unable to observe | **Behavior Descriptions**:<br>**Academically engaged** is actively or passively participating in the classroom activity. For example: writing, raising hand, answering a question, talking about a lesson, listening to the teacher, reading silently, or looking at instructional materials.<br>**Respectful** is defined as compliant and polite behavior in response to adult direction and/or interactions with peers and adults. For example: follows teacher direction, pro-social interaction with peers, positive response to adult request, verbal or physical disruption without a negative tone/connotation.<br>**Disruptive** is student action that interrupts regular school or classroom activity. For example: out of seat, fidgeting, playing with objects, acting aggressively, talking/yelling about things that are unrelated to classroom instruction. | |

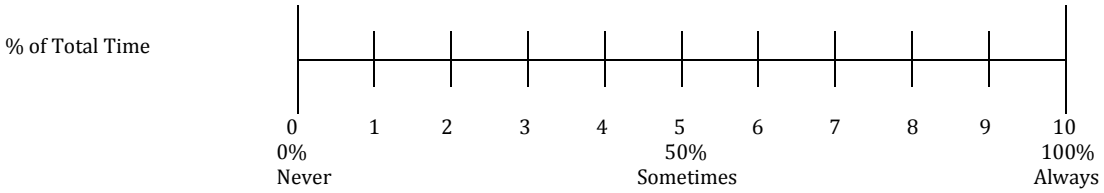
**Directions**: Place a mark along the line that best reflects the percentage of total time the student exhibited each target behavior. Note that the percentages do not need to total 100% across behaviors since some behaviors may co-occur.

Academically Engaged



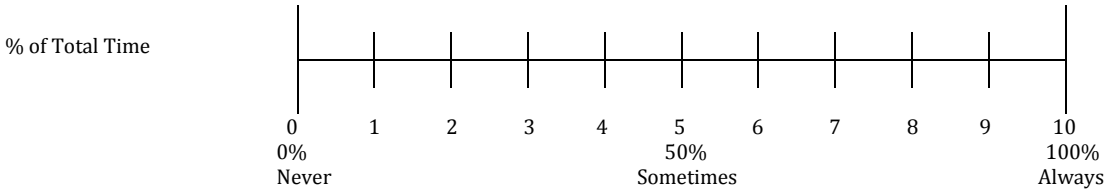

---

Respectful




---

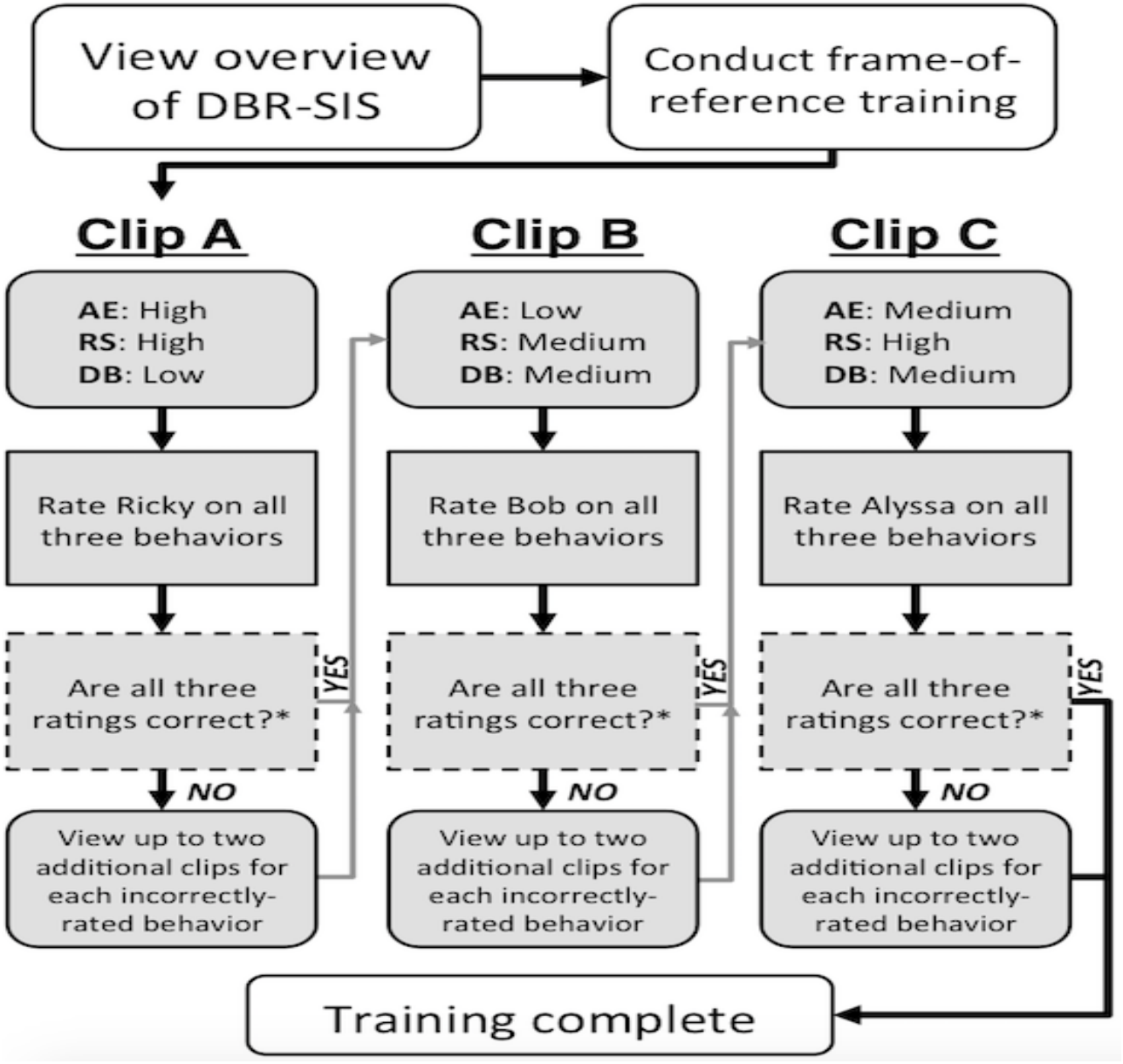
Disruptive *




- Remember that a lower score for "Disruptive" is more desirable.

Appendix B

Training Module Branching Chart (Chafouleas et al., 2015)



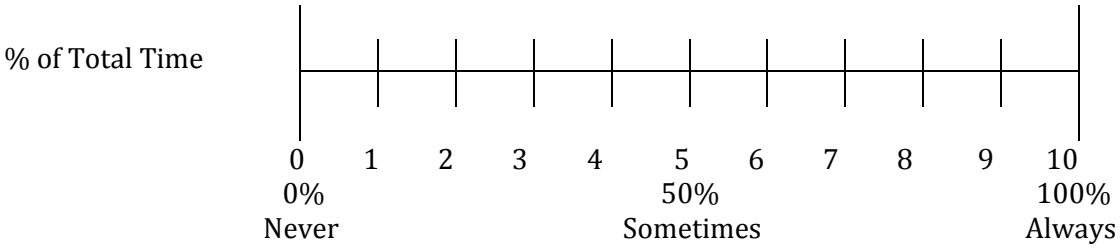

Appendix C

DBR Form For Use In the Present Study

| **Date**: _____<br>M T W Th F | **Student Name**: | **Any changes in the typical activity routine?** |
|---|---|---|
| **Observation Time**:<br>Start:_____<br>End: _____<br>☐ Check if unable to observe | **Behavior Descriptions**:<br>**Activity engagement** is actively participating in ice-skating. For example: skating on the ice or putting on equipment. Non-examples include: off the ice for any reason other than to fix equipment or standing on the ice without skating.<br>**Disruptive** is camper action that interrupts regular ice-skating activity. For example: skating in the wrong direction, playing tag on the ice, throwing snow at other individuals, or any other violation of posted rules. | |

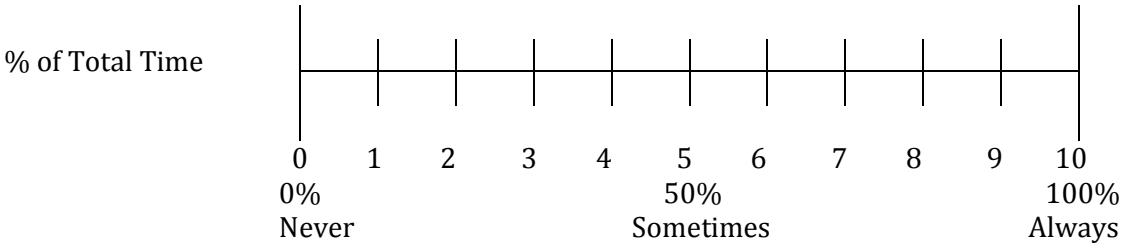
**Directions**: Place a mark along the line that best reflects the percentage of total time the student exhibited each target behavior. Note that the percentages do not need to total 100% across behaviors since some behaviors may co-occur.

Activity Engagement




---

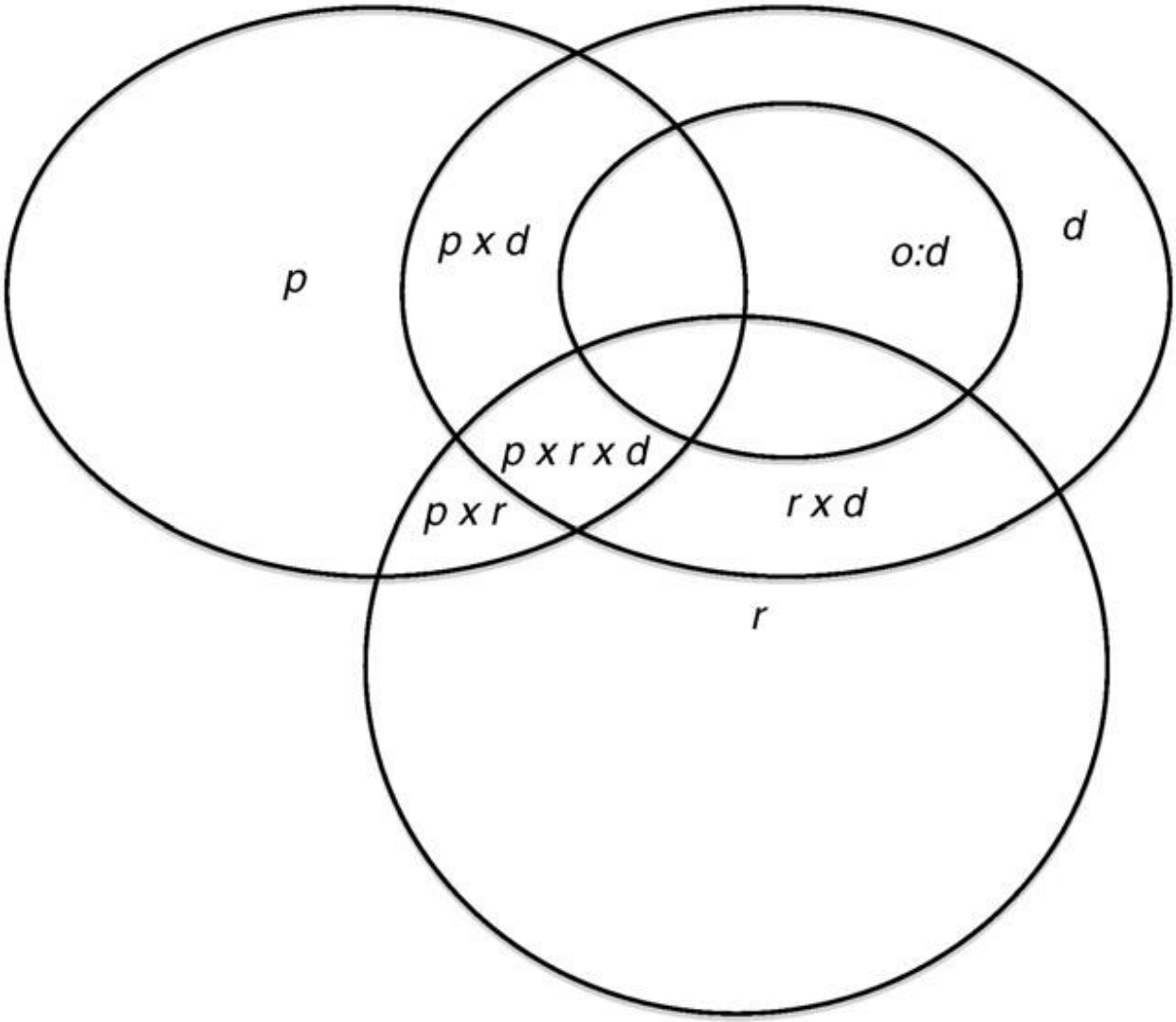
Disruptive *




- Remember that a lower score for "Disruptive" is more desirable.

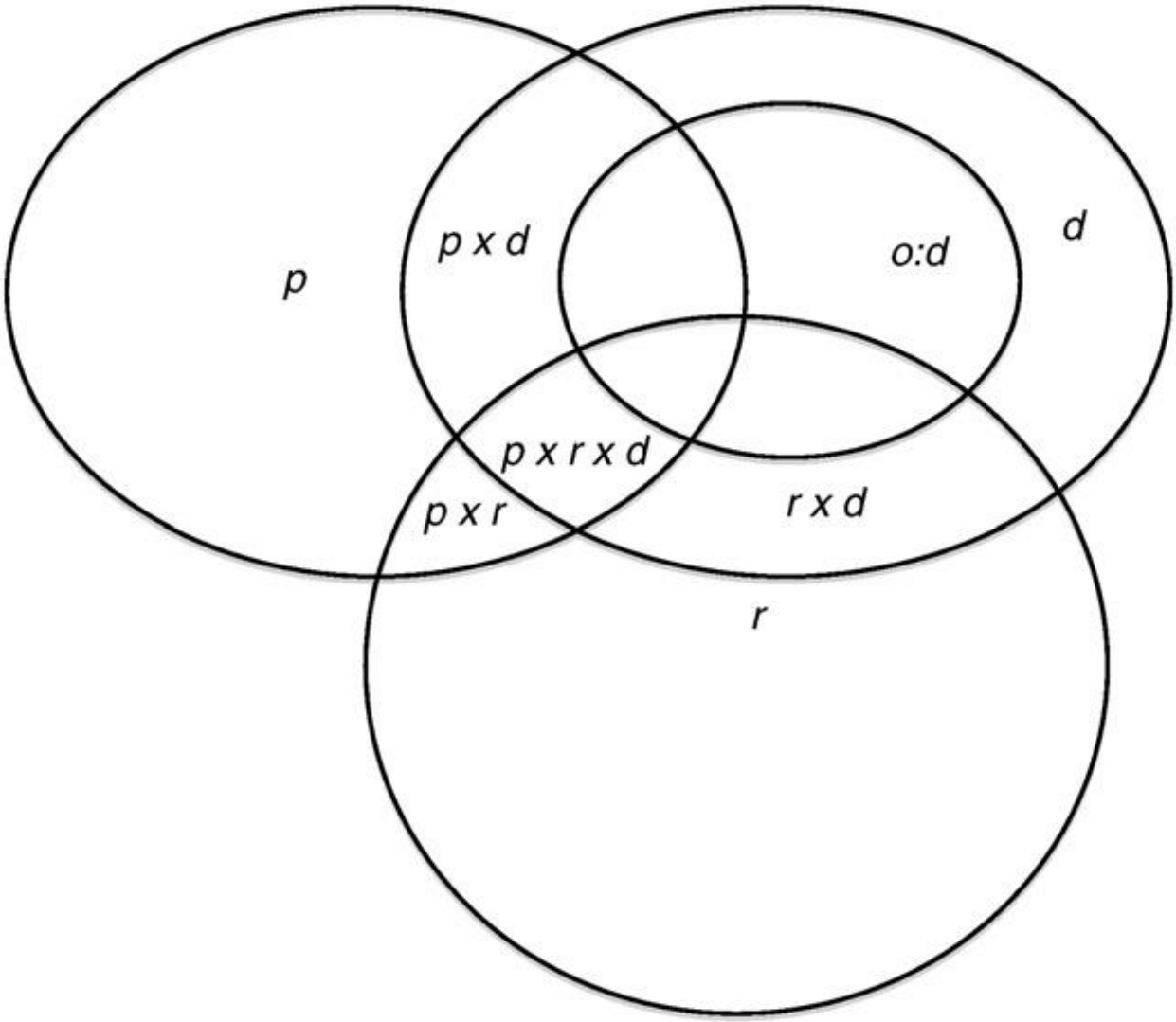
Appendix D

Step one model (two identical models for each behavior)




Appendix E

Step two model (four identical models for each behavior by rater type)




Appendix F

Step three model (eight models for each behavior type by individual rater)

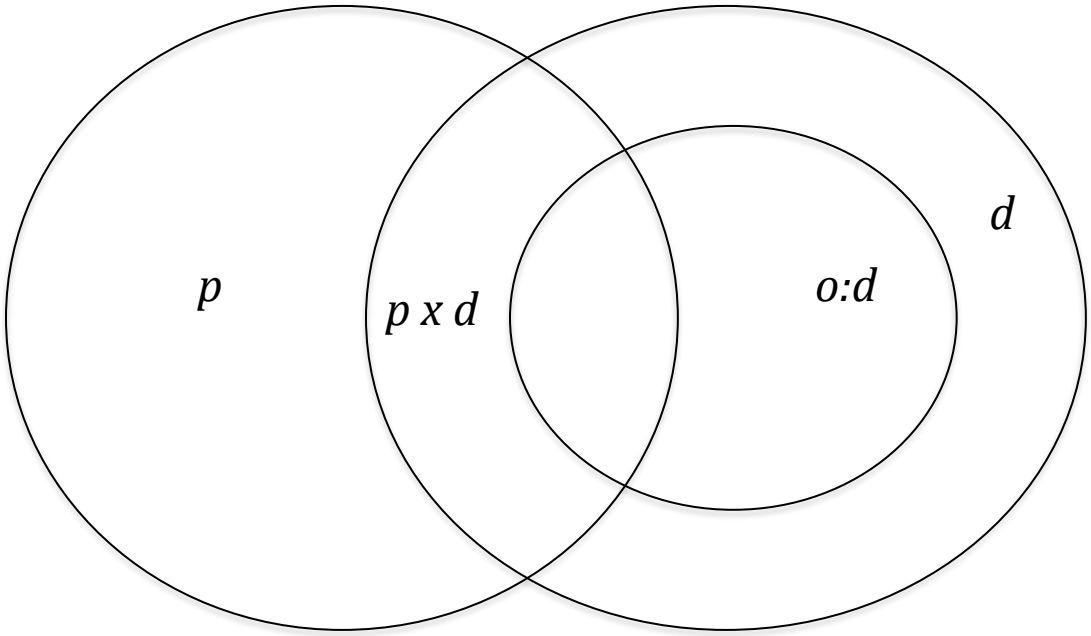


Table 1
G study results for the full model: person, rater, occasion: day, and interactions

| | AE [a] | | DB [a] | |
|---|---|---|---|---|
| | Var [b] | % Var [c] | Var [b] | % Var [c] |
| Person | 0 | 0 | 0 | 0 |
| Rater | .994 | 26% | 2.232 | 43% |
| Day | 0 | 0 | 0 | 0 |
| Occasion: Day | .08 | 2% | 0.073 | 1% |
| Person x rater | .211 | 6% | 0.128 | 2% |
| Person x day | .435 | 11% | 0.383 | 7% |
| Rater x day | .042 | 1% | 0.494 | 10% |
| Person x rater x day | .235 | 6% | 0.256 | 5% |
| Error [d] | 1.8 | 47% | 1.581 | 31% |
| Total | 3.797 | 100%[e] | 5.147 | 100%[e] |

[a] AE refers to "activity engagement" and DB refers to "disruptive behavior"

[b] Var – variance calculated using REML

[c] % Var – percentage of total variance

[d] Includes residual along with interactions involving *o:d*

[e] Values rounded to 100%

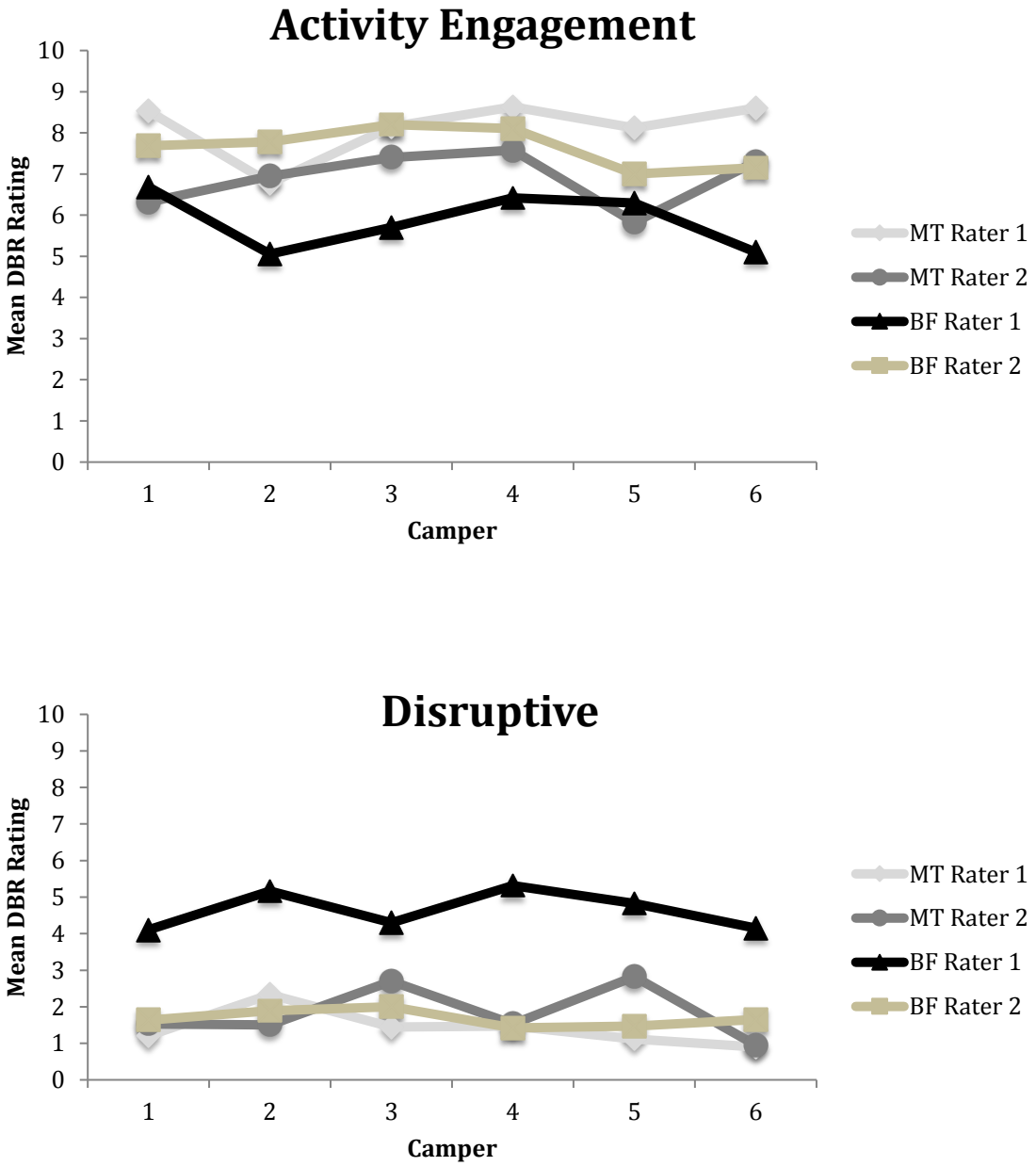
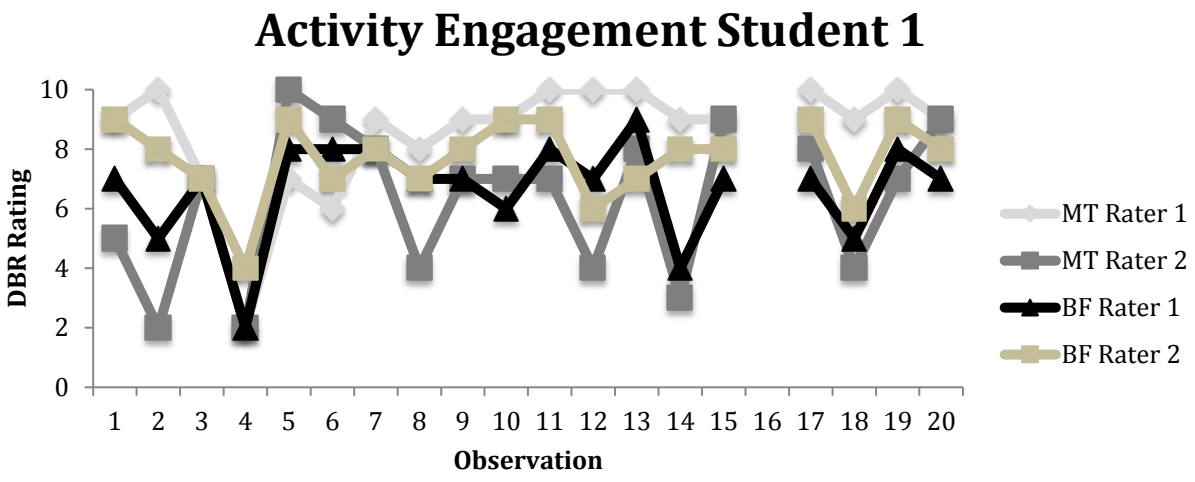
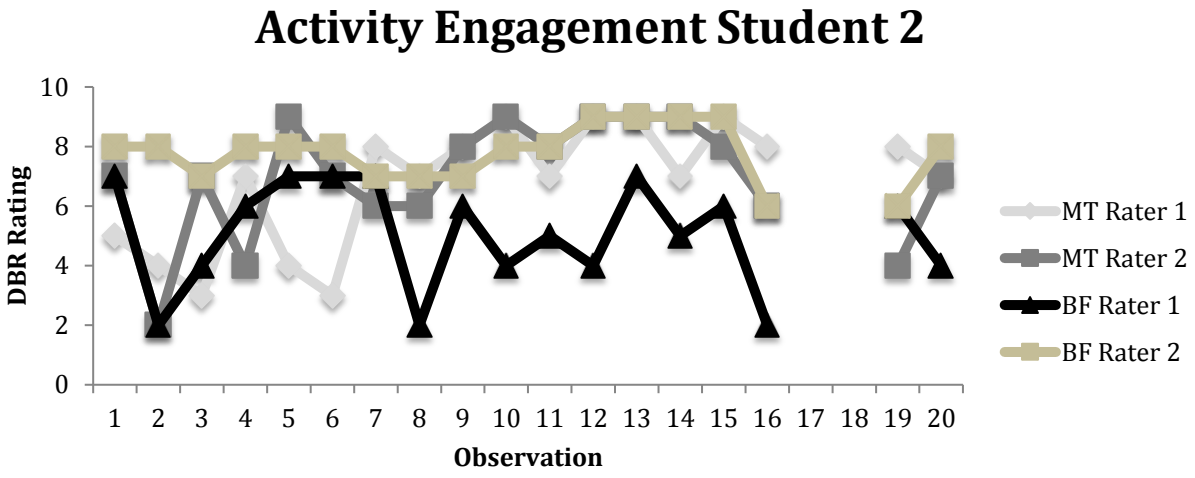
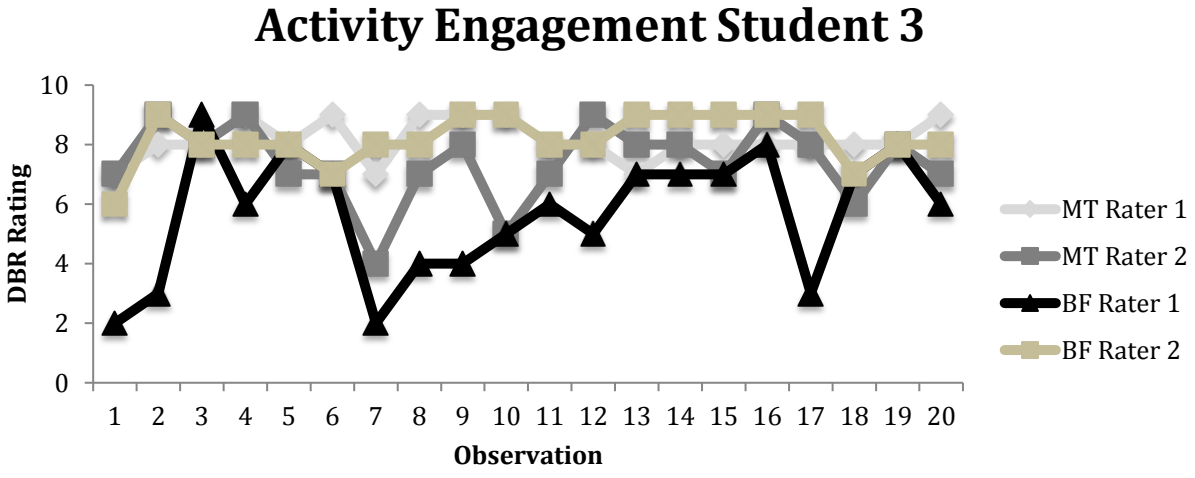



Fig. 1 Mean percentage of DBR across occasions for each camper by rater


Activity Engagement Student 1
DBR Rating
0
2
4
6
8
10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Observation
MT Rater 1
MT Rater 2
BF Rater 1
BF Rater 2


Activity Engagement Student 2
DBR Rating
0
2
4
6
8
10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Observation
MT Rater 1
MT Rater 2
BF Rater 1
BF Rater 2


Activity Engagement Student 3
DBR Rating
0
2
4
6
8
10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Observation
MT Rater 1
MT Rater 2
BF Rater 1
BF Rater 2

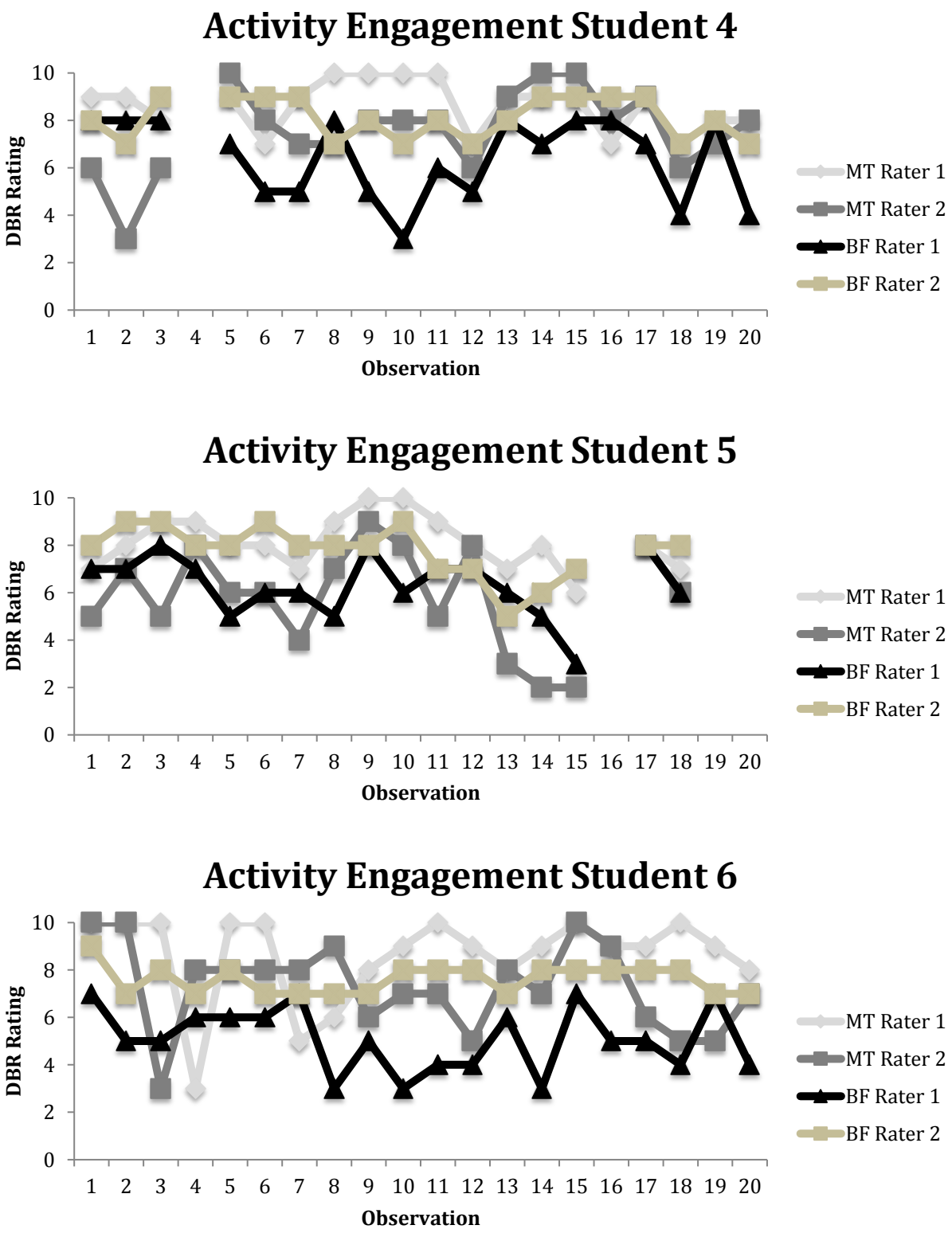
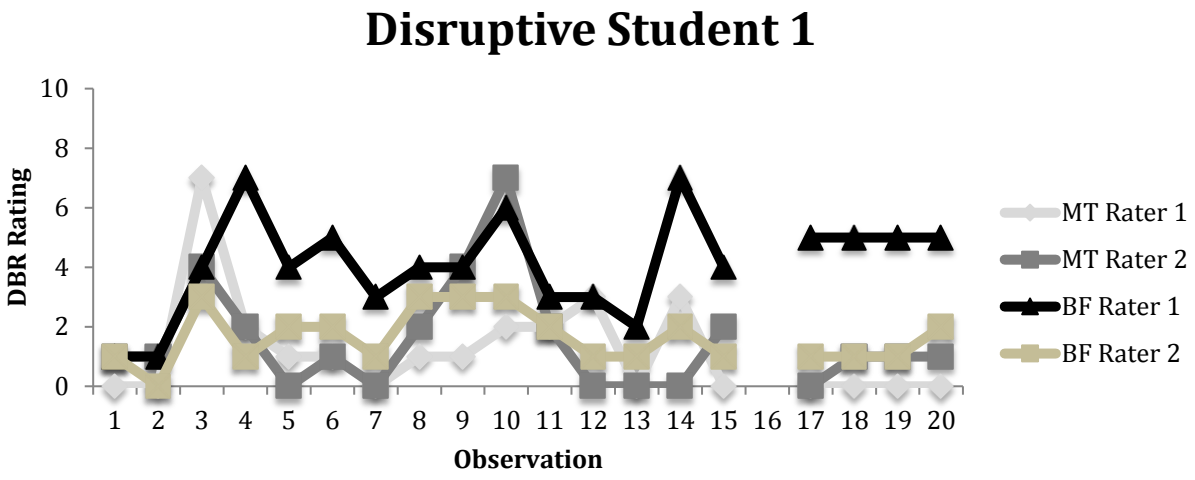
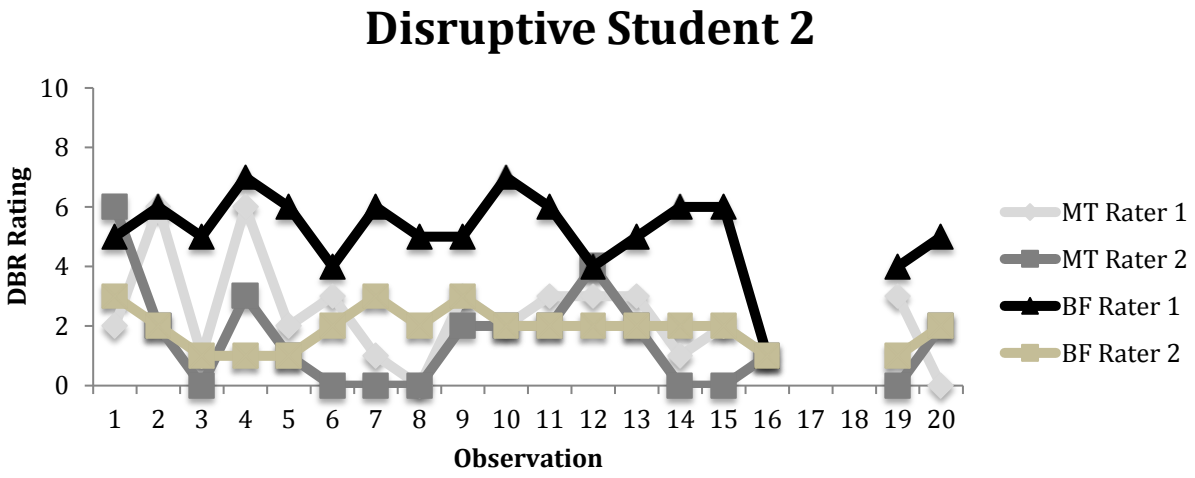
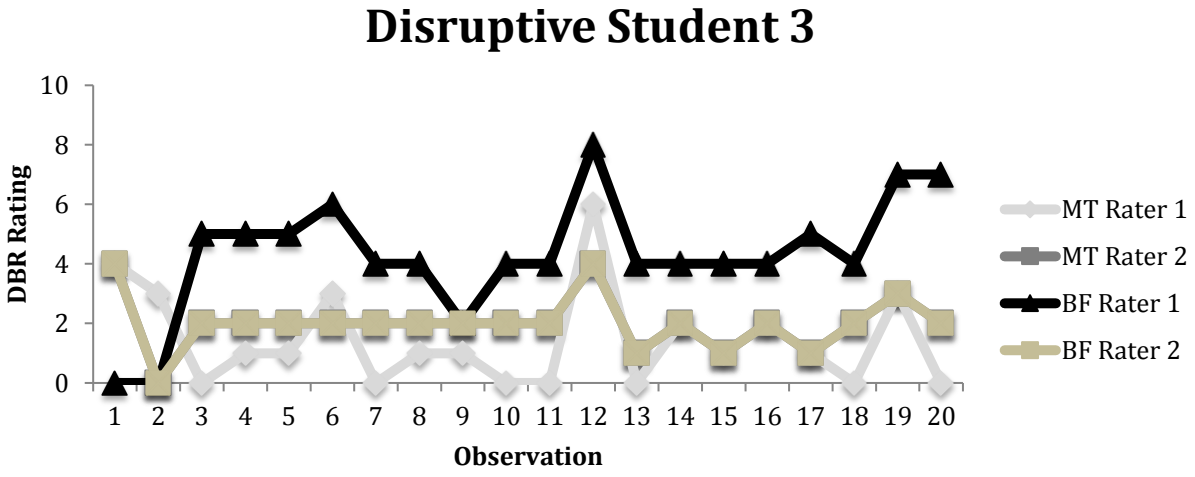



Fig. 2. Ratings of Activity Engagement for each student by rater across occasions.


Disruptive Student 1
DBR Rating
Observation
MT Rater 1
MT Rater 2
BF Rater 1
BF Rater 2


Disruptive Student 2
DBR Rating
Observation
MT Rater 1
MT Rater 2
BF Rater 1
BF Rater 2


Disruptive Student 3
DBR Rating
Observation
MT Rater 1
MT Rater 2
BF Rater 1
BF Rater 2

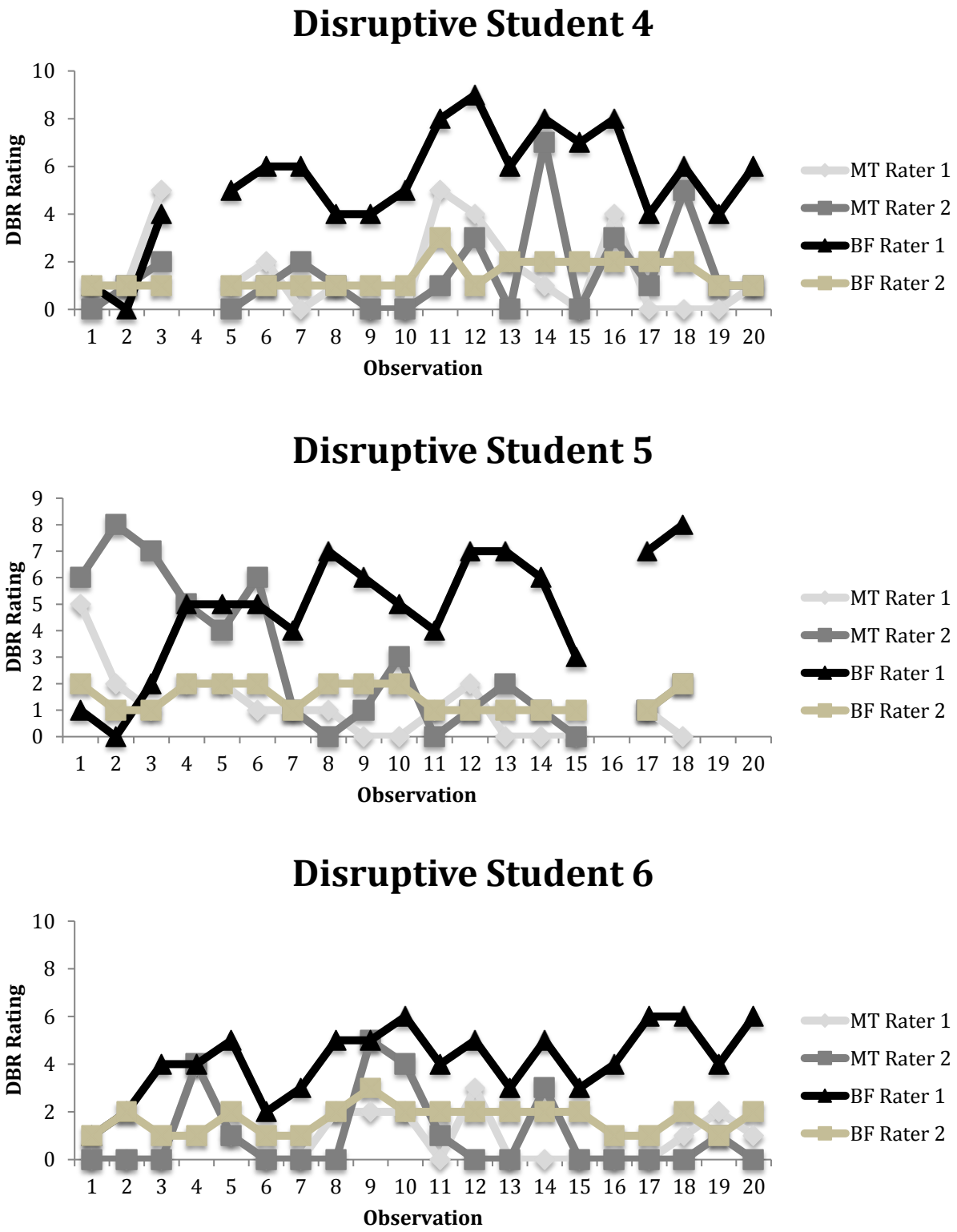



Fig. 3. Ratings of Disruptive behavior for each student by rater across occasions.

Table 2

G study results for each rater type: person, rater, occasion: day, and interactions

| | Component | Module Trained | | Brief Familiarization | |
|---|---|---|---|---|---|
| | | Var | % Var | Var | % Var |
| AE[a] | Person | 0.024 | 1% | 0.002 | 0% |
| | Rater | 0.698 | 17% | 1.927 | 48% |
| | Day | 0.058 | 1% | 0 | 0% |
| | Occasion: Day | 0.011 | 0% | 0.089 | 2% |
| | Person x rater | 0.23 | 6% | 0.158 | 4% |
| | Person x day | 0.35 | 8% | 0.294 | 7% |
| | Rater x day | 0 | 0% | 0 | 0% |
| | Person x rater x day | 0.703 | 17% | 0 | 0% |
| | Error | 2.083 | 50% | 1.544 | 38% |
| DB[a] | Person | 0 | 0% | 0 | 0% |
| | Rater | 0.027 | 0% | 4.376 | 67% |
| | Day | 0.281 | 8% | 0 | 0% |
| | Occasion: Day | 0.071 | 2% | 0.068 | 1% |
| | Person x rater | 0.211 | 6% | 0.101 | 2% |
| | Person x day | 0.685 | 19% | 0.283 | 4% |
| | Rater x day | 0.025 | 1% | 0.538 | 8% |
| | Person x rater x day | 0.156 | 4% | 0.148 | 2% |
| | Error | 2.176 | 60% | 1.017 | 16% |

Table 3

G study results for each rater: person, occasion: day, and interactions

| | Component | MT Rater 1 | | MT Rater 2 | | BF Rater 1 | | BF Rater 2 | |
|---|---|---|---|---|---|---|---|---|---|
| | | Var | % Var | Var | % Var | Var | % Var | Var | % Var |
| AE[a] | Person | 0.304 | 11% | 0.204 | 5% | 0.33 | 10% | 0.007 | 1% |
| | Day | 0.154 | 5% | 0 | 0% | 0 | 0% | 0 | 0% |
| | Occasion: day | 0 | 0% | 0 | 0% | 0.27 | 8% | 0.011 | 1% |
| | Person x day | 0.912 | 33% | 1.105 | 27% | 0.38 | 11% | 0.116 | 12% |
| | Po:d, Error | 1.413 | 51% | 2.808 | 68% | 2.239 | 70% | 0.814 | 86% |
| DB[a] | Person | 0.118 | 5% | 0.224 | 5% | 0.137 | 3% | 0.026 | 5% |
| | Day | 0.453 | 17% | 0.179 | 4% | 1.018 | 27% | 0 | 0% |
| | Occasion: day | 0 | 0% | 0 | 0% | 0.215 | 6% | 0.017 | 3% |
| | Person x day | 0.002 | 0% | 1.644 | 36% | 0.92 | 25% | 0.007 | 1% |
| | Po:d, Error | 2.031 | 78% | 2.484 | 55% | 1.444 | 39% | 0.496 | 91% |

[a] AE refers to “activity engagement” and DB refers to “disruptive behavior”

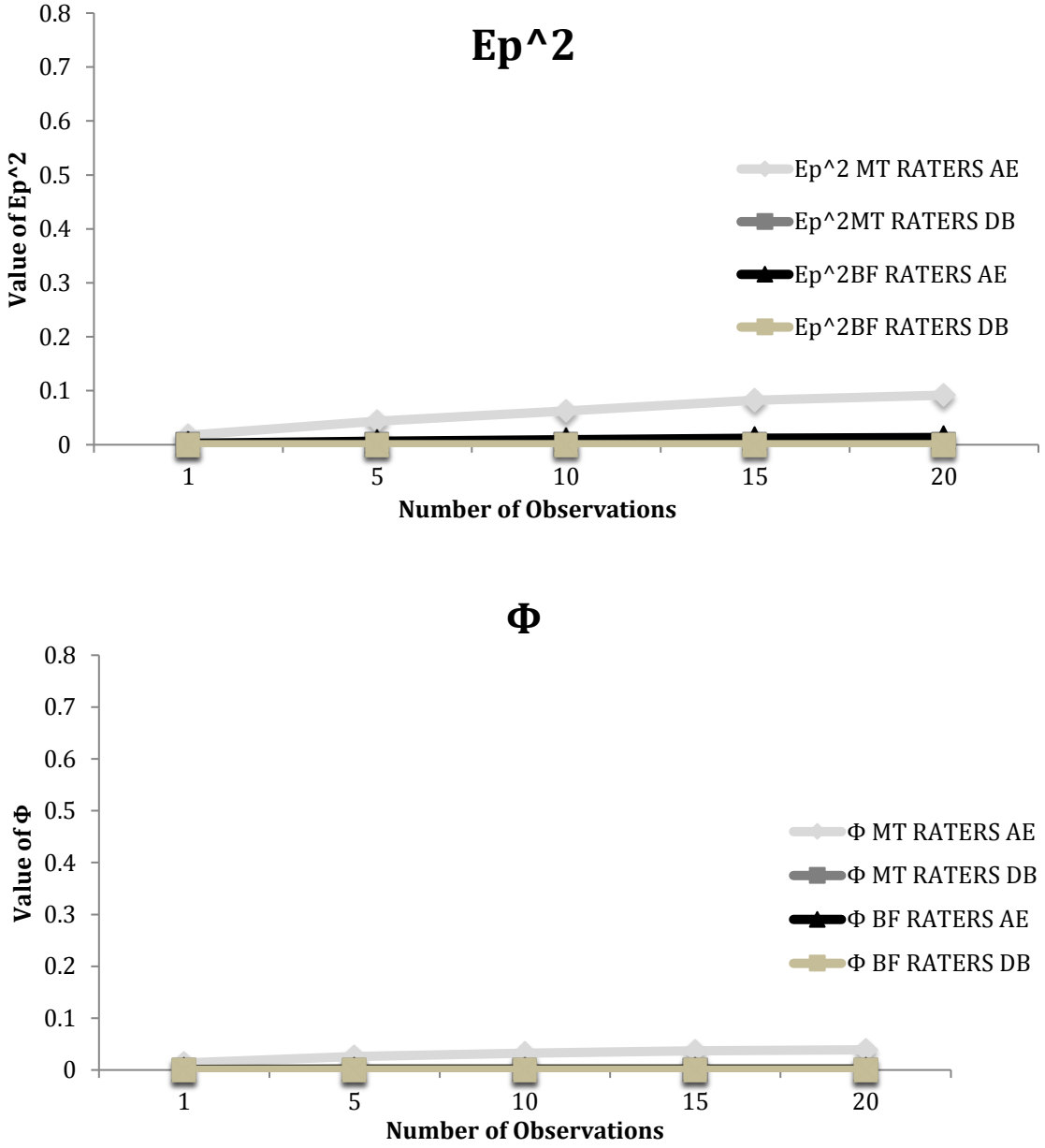



Fig. 4. Reliability like coefficients for academic engagement and disruptive behavior at selected intervals of 1, 5, 10, 15 & 20 observations for each rater type.

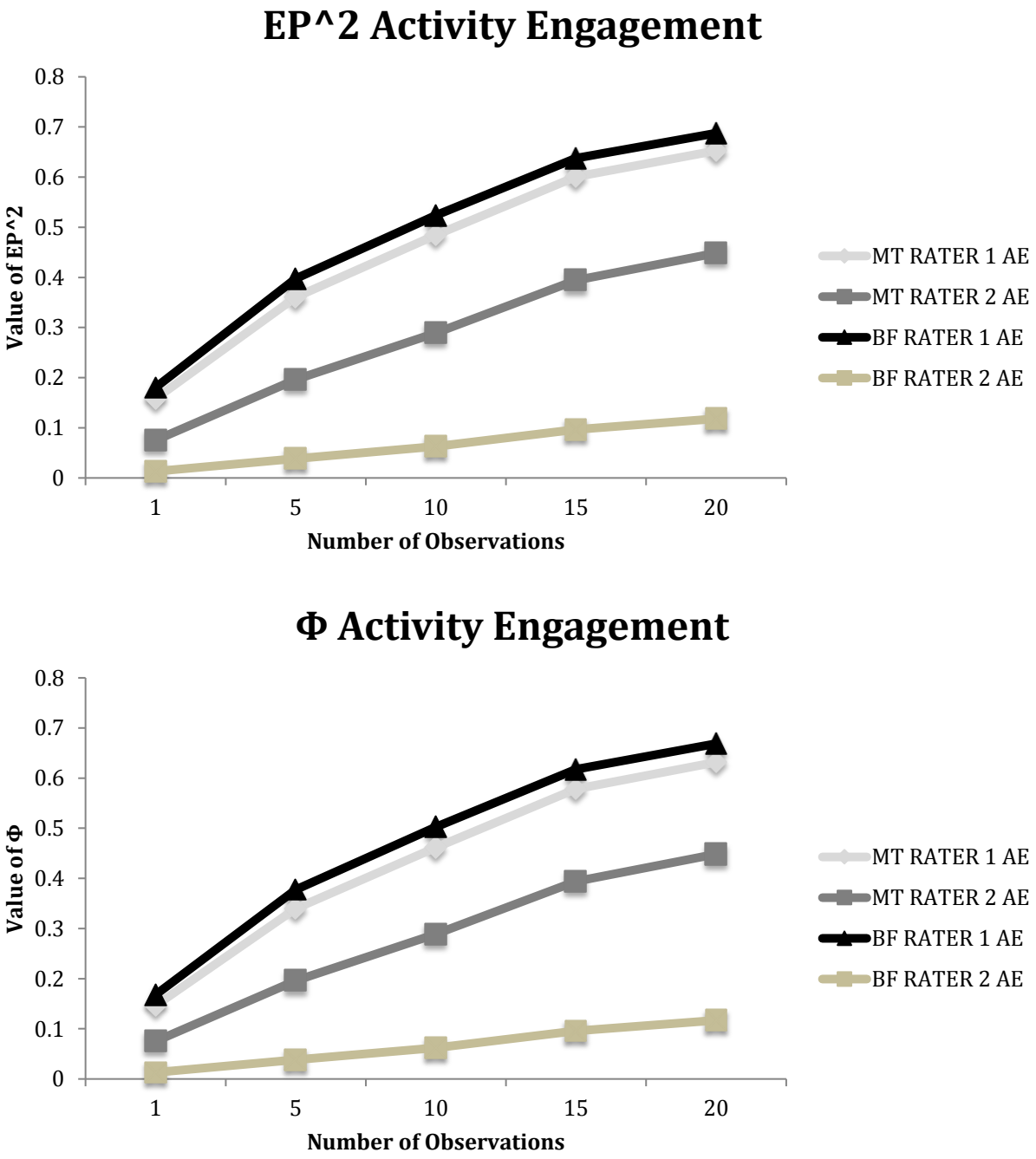



Fig. 5. Reliability like coefficients for each individual rater assessing academically engaged behavior at selected intervals of 1, 5, 10, 15 & 20 observations

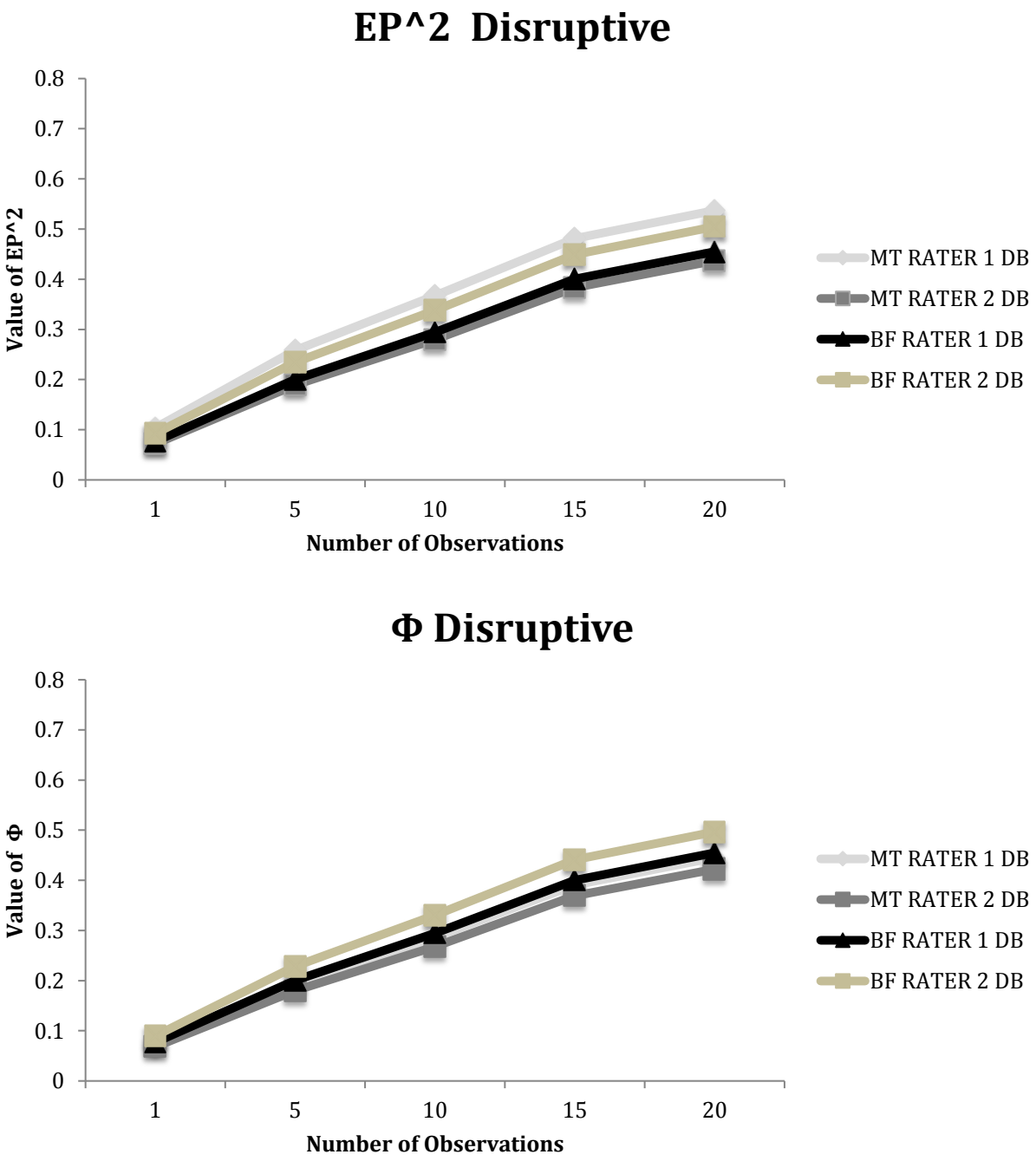



Fig. 6. Reliability like coefficients for each individual rater assessing disruptive behavior at selected intervals of 1, 5, 10, 15 & 20 observations